5-13-2022

Intelligent Data Analytics using Deep Learning for Data Science

Maria E. Presa Reyes
Florida International University, mpres029@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

INTELLIGENT DATA ANALYTICS USING DEEP LEARNING FOR DATA SCIENCE

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Maria Eugenia Presa-Reyes

2022

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Maria Eugenia Presa-Reyes, and entitled Intelligent Data Analytics using Deep Learning for Data Science, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Xudong He

Jainendra K Navlakha

Cheng-Xian Lin

Sitharama S. Iyengar

Shu-Ching Chen, Major Professor

Date of Defense: May 13, 2022

The dissertation of Maria Eugenia Presa-Reyes is approved.

Dean John L. Volakis
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

To my parents and my brother.

ACKNOWLEDGMENTS

First of all, I would like to express my utmost gratitude to my advisor Professor Shu-Ching Chen, for his invaluable guidance, encouragement, patience, and support throughout so many years of research. In addition, I would also like to thank Professor Mei-Ling Shyu of the Department of Electrical and Computer Engineering at the University of Miami (UM), professors Sitharama S. Iyengar, Jainendra K Navlakha, and Xudong He of the Knight Foundation School of Computing and Information Sciences, and Professor Cheng-Xian Lin of the Department of Mechanical and Material Engineering at FIU for the suggestions they provided. Secondly, my thanks go to the friends and colleagues from the Distributed Multimedia Information Systems (DMIS) Laboratory at FIU, the Data Mining, Database & Multimedia (DDM) Research Group at UM, and the Computational Fluids and Energy Science (CFES) Laboratory at FIU, in particular Yudong Tao, Haiman Tian, Samira Pouyanfar, Tianyi Wang, Hsin-Yu Ha, Erik Coltey, Hector Cen Zheng, Diana Machado, Raul Garcia, Yuexuan Tu, Daniel E. Martinez, Mario Jacas, Anchen Sun, Rui Ma, Christian Moreyra, Christopher Jerauld, Daniela Alonso, Beichao Hu, and Pratik Mahyawansi. Finally, but certainly not least, I am very thankful for my loving parents and my dear brother. Without their help and encouragement, I would not have been able to complete my dissertation.

This research is partially supported by NSF CNS-2125165, NSF CNS-1952089, NSF HRD-1547798, Microsoft AI for Earth - Azure Compute, Florida Public Hurricane Loss Model (FPHLM), and the Office of Fossil Energy, U.S. Department of Energy DE-FE0031904.

ABSTRACT OF THE DISSERTATION

INTELLIGENT DATA ANALYTICS USING DEEP LEARNING FOR DATA SCIENCE

by

Maria Eugenia Presa-Reyes

Florida International University, 2022

Miami, Florida

Professor Shu-Ching Chen, Major Professor

Nowadays, data science stimulates the interest of academics and practitioners because it can assist in the extraction of significant insights from massive amounts of data. From the years 2018 through 2025, the Global Datasphere is expected to rise from 33 Zettabytes to 175 Zettabytes, according to the International Data Corporation. This dissertation proposes an intelligent data analytics framework that uses deep learning to tackle several difficulties when implementing a data science application. These difficulties include dealing with high inter-class similarity, the availability and quality of hand-labeled data, and designing a feasible approach for modeling significant correlations in features gathered from various data sources. The proposed intelligent data analytics framework employs a novel strategy for improving data representation learning by incorporating supplemental data from various sources and structures. First, the research presents a multi-source fusion approach that utilizes confident learning techniques to improve the data quality from many noisy sources. Meta-learning methods based on advanced techniques such as the mixture of experts and differential evolution combine the predictive capacity of individual learners with a gating mechanism, ensuring that only the most trustworthy features or predictions are integrated to train the model. Then, a Multi-Level Convolutional Fusion is presented to train a model on the correspondence between

local-global deep feature interactions to identify easily confused samples of different classes. The convolutional fusion is further enhanced with the power of Graph Transformers, aggregating the relevant neighboring features in graph-based input data structures and achieving state-of-the-art performance on a large-scale building damage dataset. Finally, weakly-supervised strategies, noise regularization, and label propagation are proposed to train a model on sparse input labeled data, ensuring the model's robustness to errors and supporting the automatic expansion of the training set. The suggested approaches outperformed competing strategies in effectively training a model on a large-scale dataset of 500k photos, with just about 7% of the images annotated by a human. The proposed framework's capabilities have benefited various data science applications, including fluid dynamics, geometric morphometrics, building damage classification from satellite pictures, disaster scene description, and storm-surge visualization.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Introduction

The amount of data gathered is quickly increasing, with humanity doubling its data production every two years—this trend is expected to accelerate rather than slow down. As a result, Artificial Intelligence (AI) and Machine Learning (ML) tools are becoming a commodity. The importance of AI/ML technologies in our daily lives is expanding, particularly with the advent of deep learning technologies [1, 2] and the increasing demand for technologies that can help process and curate massive amounts of data to allow for the retrieval of relevant information [3, 4].

Deep learning tools and methods, particularly Convolutional Neural Networks (CNNs), have revolutionized image and video classification, significantly improving the accuracy and robustness of object detection and scene description [5]. Intelligent data analytics aims to extract valuable information from data using AI, pattern recognition, and statistics for different applications. Academics and practitioners are interested in data science because it may help obtain valuable insights from data for better decision-making [6]. For instance, high-resolution images contain many helpful fine-grained details about the scene or object captured that has been successfully leveraged in fields including but not limited to satellite land-cover classification [7], disaster damage assessment [8, 9, 10], medicine [11], and geometric morphometrics [12].

In recent years, image recognition technology has risen to become one of the most popular data science applications with the help of artificial intelligence. Image recognition often identifies specific persons, locations, and objects. For instance, Computerized Tomography (CT) scans, X-rays, and other computer-aided medical

imaging have been utilized for patient diagnosis for a long time [13]. New advancements in image recognition and data science enable physicians to better comprehend these datasets by turning them into 3D interactive models that are simple to interpret [14].

However, DNNs are still largely unexplored in many domains due to their complex nature or lack of relevant data. Most of the well-established previously proposed models also rely heavily on large amounts of well-curated data that experts have hand-labeled, which may be either expensive or unfeasible in some fields.

Photographic cameras can record and generate hundreds of high-quality data sets in a short period. Metadata and expert knowledge may be utilized with data to enhance the model's training process and supplement the features derived directly from the high-quality data collected. The supplemental information, such as expert knowledge, can further guide the deep learning training process and help tackle class imbalance and extreme inter-class similarity. Unfortunately, most expert knowledge about the data does not come with a standard arrangement, making it challenging to integrate different data modalities from various sources to train effective models.

This doctorate dissertation addresses several research challenges in data science by presenting an intelligent data analytics framework using deep learning. More specifically, the proposed study tackles several possible difficulties in a data science project, including the availability and quality of hand-labeled data and the development of a viable technique for modeling significant connections in the features derived from different data sources. The proposed framework has also been used in various applications, including fluid dynamics, geometric morphometrics, damage classification of buildings from satellite images, disaster scene description, and advanced storm-surge visualization.

The challenges addressed in this dissertation are summarized as follows:

- **Transformation and Fusion of Different Data Modalities**: Although most data come with a natural structure, such as the grid-like nature of images, most expert knowledge does not follow a similar or general pattern that can be easily modeled and fused. Similarly, some data can possess metadata that provides additional information to facilitate learning various characteristics from the data. For example, modern data collection tools may rapidly capture and produce hundreds of high-quality data. Rather than relying solely on the features obtained directly from the data captured (i.e., the visual information from an image) to train a model, metadata and expert knowledge can help enhance the training process. If expert knowledge is not readily available or is very limited for a specific application, knowledge learned by an AI model trained on a more general data benchmark can be transferred. The proposed framework is developed with the capability to adapt to different data modalities and make the transformation necessary to prepare the fused data for modeling.

- **Inter-Class Feature Similarity**: Datasets with extremely high inter-class similarity are difficult to adequately discriminate due to the minimal and often indistinguishable features that can be hard to detect. At multiple layers of the deep learning architecture, we extract and fuse fine-level patterns from the data, addressing local and global interactions of the data's features to create a more accurate prediction. A generalized approach is proposed to transform the supplemental information obtained from expert knowledge or metadata into a desirable format that can be used to identify and magnify those fine-level patterns within the input data. The transformation approach for the supplemental information depends on this information's topological structure and the specific application's research goal. We harness the power

of deep learning and its capability to maintain the prominent layout of the input data while extracting and filtering the fine-level patterns, which are then concatenated before the model makes a prediction. The proposed architecture is evaluated using a variety of large-scale, high-quality datasets from diverse domains and scenarios, demonstrating the capability of the proposed method to adapt to different data science project goals and data characteristics.

- **Limited Label Quality and Quantity**: The rate at which data is gathered and produced significantly outpaces the rate at which well-curated expert label sets are generated, which is required to train a successful model. Most well-known and previously proposed architectures for classification tasks rely on carefully curated dataset benchmarks, which may be either expensive or difficult to acquire in some instances. The proposed weak-supervision training process attempts to train a robust model considering a scenario where data labels are limited in quantity and quality. Weak supervision is a machine learning branch in which noisy, restricted, or inaccurate sources provide supervision signals for categorizing vast quantities of training data in a supervised learning environment. During training, additive zero-centered Gaussian noise is injected randomly to augment the target label, generate synthetic perturbations in the data, and reduce over-fitting [15]. Furthermore, following the natural structure (i.e., graph, sequence, or grid) assessed from the input data, a label propagation technique is proposed to help enhance the training data while the deep learning model learns from the limited label quantity.

1.2 Proposed Solutions

We propose an intelligent data analytics framework that involves gathering, organizing, and analyzing extensive data to find patterns, knowledge, and insight within the data using deep learning. Data analytics, in general, can be divided into three categories—descriptive, predictive, and prescriptive [11]. Descriptive analytics analyzes data statistically to identify what happened in the past. Descriptive analytics, for example, helps in the comprehension of a company's performance by providing context for stakeholders to interpret data. Historical data is put into a machine learning model that evaluates critical trends and patterns in predictive analytics. The model is then used to forecast what will happen next using current data. Prescriptive analytics strives to improve the accuracy of previously predicted data and provide a response to the issue of what actions should be taken, knowing what is most likely to happen in the future. In other words, prescriptive analytics presents several courses of action and describes the possible results.

This dissertation focuses on predictive data analytics by extracting insights from data and integrating additional supporting information obtained from metadata or produced by experts or other pre-trained algorithms. We offer a four-part end-to-end analytics tool incorporating data pre-processing, fusion, modeling, and visualization. We develop novel deep learning architectures to fuse and model the data pre-processed by corresponding techniques and further demonstrate advanced visualization techniques using a 3D game engine to create an immersive and interactable experience [16]. The critical components of the proposed system are evaluated using datasets for different applications, such as satellite images taken before and after a disaster event [17], a partially labeled set of pictures taken from a low-altitude

aircraft [18], and laboratory slide pictures of the right-wing of honey bees of different subspecies [12]. The analytical methods are described in full below.

1.2.1 Multi-Source Multi-Modality Information Fusion

We developed the framework to adapt to different data sources and modalities and make the proper transformations to prepare the data for fusion and modeling. Our framework harnesses the variety and valuable knowledge provided by supplemental data from other sources that can describe the contextual information found in the input data [19]. The gathered data may come in various forms, including images, tags, or concepts detected by classifiers pre-trained on more general data benchmarks. The compatibility between the data retrieved from different sources and input data is first established at the semantic level or by utilizing an apriori association that links the distinct entities. The embedding method will first represent the discrete variables as continuous vectors. Semantic similarity is applied first to measure the distance of these vectors based on the similarity of their meaning or semantic content. After determining the semantic similarity, the data is combined using the proper aggregation and normalization techniques to fuse the information for training. Other compatibility methods to match the data are also proposed based on the knowledge of the natural structure in real life. For example, spatio-temporal data can help match entities among compatible events and locations.

1.2.2 Fine-Level Pattern Modeling with Deep Feature Fusion

A fine-level pattern modeling approach using deep feature fusion is proposed to provide a robust representation capability to identify easily confused inter-class samples. A comprehensive approach is proposed for converting the additional information acquired through expert knowledge or metadata into the desired format that may be utilized to detect, filter, and magnify those fine-level patterns within the input data. The goal is to integrate the localized features extracted from earlier layers to provide a more accurate prediction, especially for overlapping classes. We further use deep learning's capability to preserve the dominant structure of the input data, especially in the case of CNN, to extract the fine-level patterns from different layers at the same region of interest. Depending on the project goal, these local features may subsequently be concatenated with global features produced by later layers of the network that include more broad information about the input data.

1.2.3 Weakly-Supervised Training

We propose an end-to-end weakly-supervised model [19, 15] where the assumption is that the label set is limited and might include instances of mislabels and errors. During training, random additive zero-centered Gaussian noise is introduced to create synthetic perturbations in the target label, augment the data, and minimize overfitting. These minor perturbations are intended to assist in training a model that is resistant to the noise present in real-world data, such as coordinating position inaccuracies and crowd-sourced mislabels or biases. Our proposed framework combines a weakly-supervised deep learning method with a novel label propagation model. The label propagation technique improves the training data by assigning labels to

previously unlabeled data to enhance the model's contextual awareness [19]. In the TRECVID2020 Disaster Scene Description and Indexing (DSDI) Challenge [20], our method achieved the top score out of all submitted runs regardless of training type, showing its better capabilities compared to other proposed approaches.

1.2.4 3D Advanced Visualization

Three-dimensional representations of real-world locations have been often used in computer and mobile phone games and virtual globes to give players unique and immersive experiences. These sophisticated visualization tools are well-known and have been successfully utilized to study complex data from multiple sources interactively. This dissertation delves into a case study of a storm-surge simulation utilizing real-world Geographic Information System (GIS) data and estimated damages to create an automated virtual environment.

1.3 Contributions

The following are the dissertation's main contributions:

- We built the intelligent data analytics framework to be adaptable to various data sources and modalities and perform the necessary transformations to prepare the data for fusion and modeling. Our technique capitalizes on the diversity and importance of supplemental data from various sources that may help explain the contextual information included in the input data. Images, tags, and ideas identified by classifiers pre-trained on more broad data standards may be among the data collected.

- A deep neural network aided by scientific knowledge is trained to predict the particle drag force coefficient from single-particle experimental data. Aside from the standard sphericity and Reynolds number, the proposed technique includes additional literature-supported parameters such as density, aspect, lengthwise and crosswise sphericities. Despite the limited data and the volatility in each single-particle experiment, model regularization and meta-learning assist train a broader and more reliable drag model.
- Fine-level pattern modeling is proposed to capture the fine-grained details at multiple levels of the architecture to address the inter-class similarity. The local and global features captured from the image are then fused before making a final prediction regarding the level of damage that the target building withstood. This processing considers challenges such as intra-class variations and inter-class similarities that can be found in the data.
- A two-stream deep feature fusion network is intended to take two picture patches as input and estimate the degree of damage to the building in the patch's center. We demonstrate how our proposed work improves the model's performance by first preprocessing the input data in a unique way that reduces uncertainty and improves model performance and then applying a new network configuration that focuses on a fusion technique that is more advanced than simply concatenating the deep features from each stream.
- A weakly-supervised training method is proposed to acquire knowledge from noisy, limited, and imprecise labels found in low-altitude disaster imagery. Multi-modality information inferred and the sequence-based information obtained from the low-altitude images are leveraged to augment the training dataset. The proposed method is evaluated on the LADI dataset as one submitted solution in the TRECVID2020 [20] Disaster Scene Description and

Indexing (DSDI) Challenge. Our proposed solution achieved the best score among all the participants.

- Multi-source weak supervision fusion is used to train an imbalanced dataset with noisy labels. Using Confident Learning, we reduce noise while improving label quality. We aggregate relevant predictions from models trained on large-scale visual datasets to improve performance on underrepresented target features. Our method outperformed all other submitted solutions in the TRECVID2021 [21] DSDI Challenge, regardless of the training data.
- The proposed framework's capabilities have benefitted a variety of data science applications from different fields, including fluid dynamics, geometric morphometrics, building damage classification from satellite images, disaster scene description, and storm-surge visualization.

1.4 Scope and Limitations

The following assumptions and restrictions apply to the proposed framework.

- The proposed intelligent data analytics platform uses pre-trained models, including object identification and segmentation, to generate additional information that can help enhance the training data, especially in the case of weak-supervision. However, the scope of this dissertation does not include object identification or semantic segmentation. Well-established approaches in localization and segmentation are used to demonstrate how such techniques may function with the suggested intelligent data analysis platform.
- Without loss of generality, the proposed intelligent data analytics platform is mainly assessed on image information along with some supplemental data,

which comprises metadata, expert knowledge, and the predictions made by pre-trained classifiers. However, it is possible to expand the proposed techniques to encompass additional data types, including multi-source and multi-modal data fusion, fine-level pattern modeling, and weakly-supervised training.

- Several of the parameters are experimentally determined. The best training performance is utilized as a measurement to derive the optimal hyper-parameters for the components featured in the proposed framework.

1.5 Outline

The following is the structure of this dissertation: The literature overview for multi-source and multi-modal data fusion, fine-level pattern modeling, and weakly-supervised training is presented in Chapter 2. The overall framework and its proposed components are introduced in Chapter 3. Chapter 4 describes many solutions to harnessing the knowledge from different data modalities and sources through fusion at both the data-level and feature-level. The fine-level pattern modeling component is introduced in Chapter 5 to address typical problems with high inter-class similarity. Chapter fc6:weak demonstrates a weakly-supervised end-to-end model that uses data from restricted label sets to incorporate perturbations in the data and augment the training. Furthermore, a 3D visualization is described in Chapter 7. Finally, Chapter 8 concludes with recommendations for further research.

CHAPTER 2

RELATED WORK

The related work in multi-source and multi-modal data fusion, fine-level pattern modeling, and weakly-supervised training is discussed in more depth in this chapter.

2.1 Multi-Source and Multi-Modality Information Fusion

Multi-modal data has emerged as a prominent area of big data, with each modality encoding a separate property of data objects. Different modalities are often complementary, prompting research into integrating multi-modal feature spaces to characterize data objects better [22, 23].

Data fusion combines information from various modalities and sources to increase prediction accuracy. However, the significant data variations in these multi-source and multi-modal datasets provide major technical challenges for machine learning and deep learning algorithms. To address this problem, scientists have developed several fusion approaches that combine data from various modalities, such as image, text, and video, to improve the model's dependability, resilience, and accuracy. Furthermore, earlier information fusion research has often required data from all modalities to be readily accessible for each training data instance. Because missing modalities exist in real-world applications, this situation substantially limits the applicability of information fusion approaches.

Data-level fusion, feature-level fusion, and decision-level fusion are three kinds of recently suggested fusion methods [24]. By combining the separate raw multi-modality data from many sources, data-level fusion algorithms make predictions before any modeling is done. Feature-level fusion algorithms anticipate the fusing of features from several modalities. Instead of combining low-level raw sensor data into data-level fusion or functionality abstracted from raw sensor data in feature-level

fusion, decision-level fusion procedures fuse high-level decisions based on individual sensor data. In the following subsections, we'll go through each degree of fusion in detail.

2.1.1 Data-level Fusion

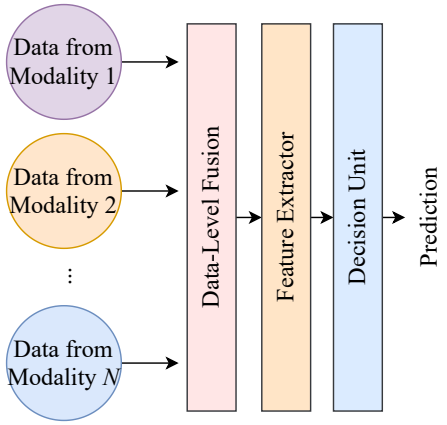


Figure 2.1: Data-level fusion.

Data fusion, also known as information fusion, combines data from many sources to build more sophisticated models or get a better understanding of a particular project. It usually entails gathering and combining data for fundamental analytical objectives on a specific issue. For example, scientists may utilize data fusion to integrate physical tracking with environmental data in a study endeavor. To generate a better picture, marketers may combine consumer identity data with purchase history and other data obtained at brick and mortar stores. Data fusion has been presented to ensure data reliability when using sensor-generated data [25]. Hard and soft data are two distinct categories of produced data that are taken into account during data-level fusion. Their bias, observational levels, consistency, and conclusions are the most notable differences amongst them [26]. Most information

fusion research has focused on hard data and very little on soft data [27]. Hard data is generated by equipment and applications such as smartphones, computers, sensors, water meters, traffic surveillance systems, phone call logs, and bank transaction records [28]. Soft data is defined as human intelligence, which contains views, recommendations, interpretations, inconsistencies, and uncertainties.

Hard Data Fusion

The literature on the fusion of traditional data supplied by nonhuman (hard) sensors is extensive and well-established [29, 30]. It is relatively common knowledge that various channels or sensors aid in identifying an item in many research fields, such as remote sensing, pattern recognition, and multimedia applications. Hard multisensor data fusion is widely utilized in intelligent environments to combine data from disparate sensors. On the other hand, sensors offer noisy and unreliable data, posing a significant challenge to researchers. Many studies [31, 32, 33] have concentrated on developing new deep neural network architectures for the identification of human activities based on various sensor data sources. Some of the proposed architecture encodes the time series of sensor data like images and leverages these transformed images to retain the necessary features for human recognition of human activity.

Soft Data Fusion

Soft data can range from crowdsourced 'soft' reports from human observers to opinion reviews on an online market website. Opinions, ideas, interpretations, inconsistencies, and ambiguities abound with soft data. Sentiment analysis on microblogs based on many user comments and messages is a fascinating field of data mining and Natural Language Processing (NLP) [34]. Several techniques and algorithms for conducting sentiment analysis on Twitter have recently been developed. Several

recommended classification and pure NLP-based methods predict sentiment orientation in various ways. NLP, data transformation, semantic data structures, soft data association, and graph matching are among the topics mentioned by [35] and [36].

Soft-Hard Data Fusion

This human-observed data has a lot to offer to gain complete situational awareness, especially in areas like intelligent analytics, where tiny linkages and interactions are difficult to detect with physical sensors. Gross *et al.* [35] present a collection of algorithms for effectively fusing hard sensor data with soft social network data (tweets) to forecast events like revolutions and terrorist acts.

2.1.2 Feature-level Fusion

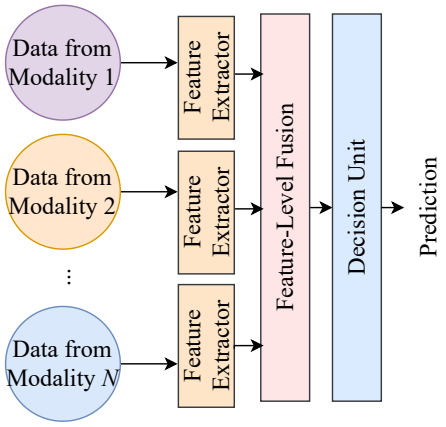


Figure 2.2: Feature-level fusion

Feature fusion methods are concerned with selecting and combining features to eliminate duplicate or unnecessary characteristics [37]. If a feature has a weak correlation with the class information, it is considered irrelevant. After extracting the feature from each dataset or data modality, the definitive collection of features is

combined to create a more useful feature set, fed into a classifier to produce the final output. The easiest method to fuse multi-modality features is to sum or concatenate feature maps from two or more levels.

Feed-forward Feature Fusion

Neural networks are a universal function whose performance increases in accuracy as the number of layers rises. However, there is a limit to how many layers can improve model accuracy. If the number of layers is increased indefinitely, performance will ultimately saturate and degrade, which is referred to as the degradation issue in practice. Feedforward feature fusion is used by ResNet [38] and DenseNet [39, 40] to offer a viable solution. Adding more layers to a network can also result in a problem known as degradation, where accuracy gets saturated as the depth of the network increases, leading to a higher training error. Deep residual networks (ResNet) [38] address these problems by breaking down a deep neural network into smaller sections connected through skip connections which form a bigger network. ResNets learn the residual representation functions instead of the direct signal representations.

DenseNets are convolutional neural networks with dense connections between layers. Each layer takes incoming inputs from all previous levels and passes on its feature maps to all following layers to maintain the feed-forward nature. A Dense Network connects all levels directly to one another. Although DenseNet outperforms ResNet in specific ways, issues like high GPU memory consumption during training significantly restrict its use. It is essential to improve its training efficiency for future applications.

Multi-modality Feature Fusion

Earlier works in multi-modality feature fusion used simple methods such as concatenation [41, 42], while more recent research suggests more complex fusion methodologies. In 2016, Feichtenhofer *et al.* [43] was one of the earliest research works to introduce a novel and unique study that investigated a variety of methods for spatially and temporally combining CNN architectures to use the spatio-temporal information extracted from different modalities effectively. Poria *et al.* [44] propose a Long Short-Term Memory (LSTM) model that takes as input a video sequence of statements and extracts contextual uni-modal and multi-modal characteristics by modeling the relationships between the input statements. Zadeh *et al.* [45] present a Tensor Fusion Network to explicitly combine uni-modal, bi-modal, and tri-modal interactions. Liu *et al.* [46] propose a Low-rank Fusion Network to overcome the disadvantage of a high number of parameters by using a low-rank factor. Chen *et al.* [47] provide a Gated Multimodal Embedding model for learning to filter noisy or conflicting modalities.

2.1.3 Decision-level Fusion

Each classifier in decision level fusion applies a threshold to the match score and produces a decision or predictive output. The results of several classifiers are then combined to create the final judgment. Bayesian inference [48] and Dempster-Shafer [49] fusion are two classic decision-level fusion methods. Bayesian Inference has shown effectiveness when previous information regarding sensor reports is accessible and provided. On the other hand, Dempster-Shafer fusion was proposed to explicitly remove such a limitation on information beforehand at the expense of a significant increase in computing complexity. As an evidence-based fusion technique, Demp-

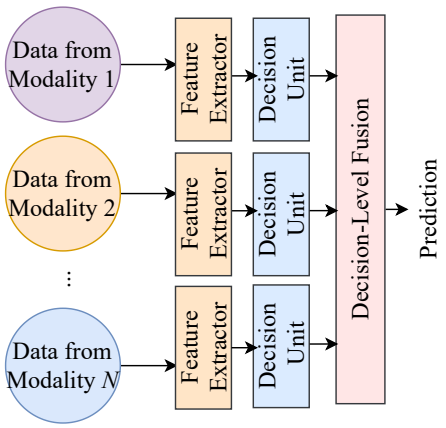


Figure 2.3: Decision-level fusion

ster–Shafer evidence theory is one of the most successful methods for data fusion. It is helpful for modeling and processing ambiguous information to convert contradictory facts into decision-making outcomes.

2.2 Fine-Level Pattern Modeling with Deep Feature Fusion

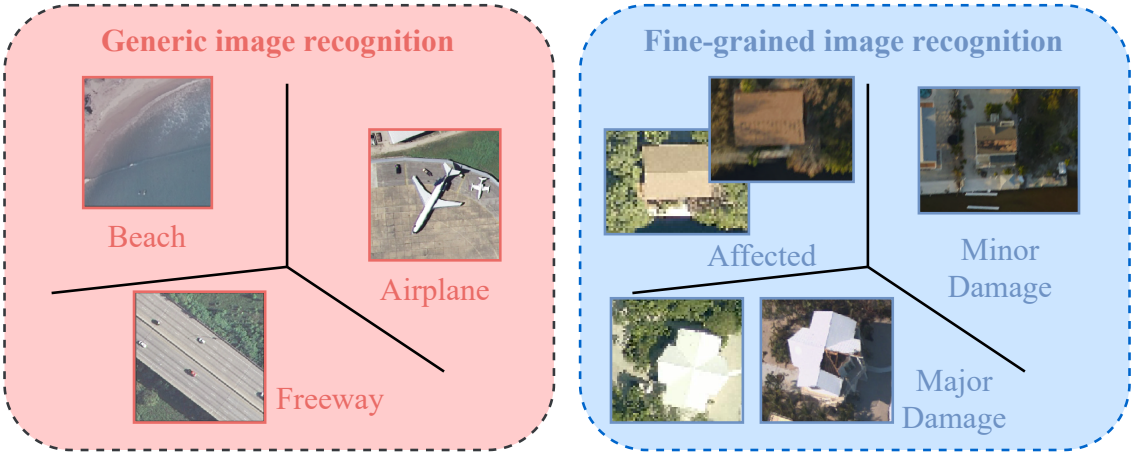


Figure 2.4: This figure compares a generic image analysis and fine-grained image analysis scenario, using the satellite image recognition task as an example.

Fine-Level pattern modeling seeks to distinguish fine features among visually similar groups. Class overlap, also known as inter-class similarity, occurs when large regions in the data are from two or more highly similar classes. Because of the significant inter-class resemblance, differentiating between two or more types is difficult, if not impossible. Other issues, such as class imbalance [50], are exacerbated by overlapping features between minority and majority classes. It is essential to mine fine-level patterns from data to find more fine-grained characteristics that may successfully aid the model in distinguishing between various target categories.

Some researchers have recently shown that features derived from a few key portions of the input data may infer more discriminate information than those retrieved from the entire data [51, 52, 53]. In the case of an image, local features describe picture patches (often composed of a small group of pixels), while global features explain the whole image [54]. For instance, Fang *et al.* [55] propose a coarse-to-fine CNN for fine-grained vehicle model identification. CNN feature maps are used to identify the most discriminant components automatically. Applications of fine-level pattern modeling range from car make and model recognition [56], plant disease classification [57], person identification [58], and dog breed recognition [59].

A relevant feature (global or local) includes discriminating information that allows one object to be distinguished from others. However, global features have a few drawbacks, such as susceptibility to noise, changes in lighting, scaling, and the inability to identify the image's essential characteristics [60]. Local features overcome their disadvantages, which encode local information to get better picture details such as spots of interest.

Deng *et al.* [37] developed a human-in-the-loop classification game that demonstrated the significance of discriminative components in fine-grained tasks. In 2013, Berg & Belhumeur [61] was the first to show that part-based one-versus-one dis-

criminative features may be helpful. Krause *et al.* [62] further expanded on this work by presenting a co-segmentation-based approach for generating discriminative parts without part annotations. Wang *et al.* [63] recently proved that adding geometric restrictions between triplets of discriminative features improves the usage of parts. Moreover, Bai *et al.* [64] suggest using the intra-class variance in triplet network metric learning to enhance recognition performance. Recently, Lin *et al.* [65] presented a network with two streams of appearance models and a bilinear pooling layer, claiming that manually selecting discriminating sections of the data was inefficient. The bilinear pooling method could implicitly explore optimum portions. The performance of bilinear CNNs in fine-grained recognition is superb; however, bilinear features' very high dimensionality renders them unsuitable for actual applications, particularly large-scale ones.

The fine-grained categorization issue has been further addressed via the use of networks that capture multi-scale features. One of the most effective techniques in multi-scale feature modeling is spatial pyramid pooling [66, 67, 68], which is an extension of the Bag-of-Words (BoW) model. It divides the picture into finer to coarser divisions and aggregates local characteristics within them. ASSP (Atrous Spatial Pyramid Pooling) [69] is a semantic segmentation module that resamples a given feature layer at various rates before convolution. This entails probing the original picture with several filters with complementing fields of view, collecting objects, and essential visual context of different sizes. Rather than resampling features, the mapping is accomplished via several parallel atrous convolutional layers with varying sampling rates.

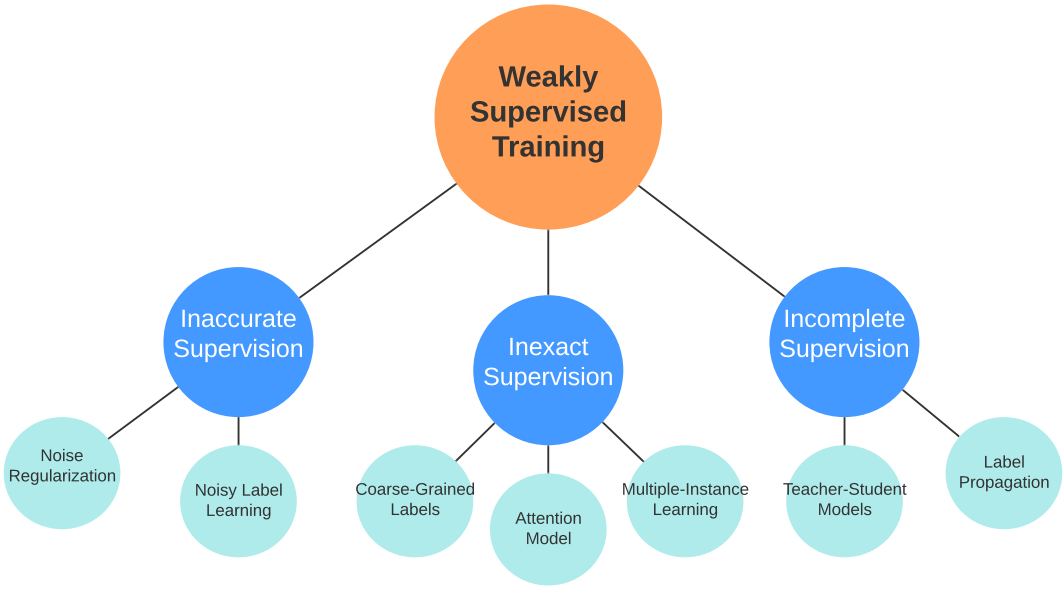


Figure 2.5: Three well-known characteristics of weakly-supervised training

2.3 Weakly-Supervised Training

Deep learning models have essentially eliminated the job of hand-crafted feature engineering owing to their remarkable ability to learn representations across numerous domains and activities autonomously. However, most of these deep learning models are complete black boxes, with no control for the ordinary developer other than labeling massive training sets and adjusting the network architecture. In many ways, deep learning models are the polar opposite of traditional expert systems' rigid but readily controlled rules—they are flexible but difficult to govern. Researchers are beginning to consider harnessing domain experience or task expertise to create more robust modern deep learning models.

A good number of currently-developed AI systems are supplied with annotated data in the case of supervised learning. However, when we work with larger models, labeling all data gets more challenging. Furthermore, there is just insufficient

labeled data for a few jobs that are only now beginning to consider AI systems to extract useful information from raw data and facilitate decision-making. For these reasons, researchers and practitioners are increasingly using lesser types of supervision, such as creating training data heuristically using external knowledge bases, patterns, or other classifiers. Essentially, these are all methods for creating training data programmatically. Using higher-level or noisier feedback from subject matter experts is what poor supervision entails.

There are three kinds of weak-supervision in general [70]. The first is incomplete supervision, in which just a small portion of training data is identified while the rest is left unlabeled. The second kind of supervision is imprecise supervision, in which only coarse-grained labels are provided. The third kind is erroneous supervision. The labels provided are not always correct—which may occur when the image annotator is inattentive or exhausted or when some images are challenging to categorize.

Weak supervision is often used after obtaining a small, high-quality dataset. Denoising and training approaches modularize the process by first modifying label confidences using just rules correlations and then training a model with the modified labels [71]. Well-known weakly-supervised tools include Knodle and WeakClust. Knodle [72] is a software framework that separates weak data annotations from deep learning models and considers them independent, modular components. It provides fine-grained information to the training process, such as data set characteristics, heuristic rule matches, or model components eventually utilized for prediction. WeakClust [73] propagates a partly labeled cluster's labels to the whole cluster, while WeakCert provides the trained classifier's predicted labels.

2.3.1 Incomplete Supervision

Incomplete supervision refers to the scenario in which we are provided a small quantity of labeled data that is inadequate to train a decent learner, despite the availability of large amounts of unlabeled data. In the absence of full supervision, either active learning [74] or semi-supervised learning [75] may be used, depending on the level of human involvement—also coined as human-in-the-loop. Active learning involves a domain expert in obtaining labels for unlabeled samples on the premise that the number of queries only dictates the cost of labeling. Hence, the aim is to minimize the number of queries. On the other hand, semi-supervised learning uses generative models [76], label propagation [77, 78], and transductive support vector machines (TSVM) [79] to exploit unlabeled data without requiring a human-in-the-loop.

There are two fundamental assumptions in semi-supervised learning: the cluster assumption [80] and the manifold assumption [81], which are related to data distribution. The cluster assumption asserts that data has a built-in cluster structure and that instances in the same cluster have the same class label. The manifold assumption holds that data is distributed on a manifold and that adjacent examples will have comparable outcomes. Both assumptions are based on the idea that equivalent data points should provide similar results, while unlabeled data may help reveal which data points are related.

Label Propagation

Because supervised learning requires a large quantity of labeled data, adequate labeling is one of the most critical aspects of machine learning success. Label propagation or disagreement-based techniques create several models and allow students to educate one another on how to exploit unlabeled data; it's a hybrid of semi-supervised

and active learning. Previous research has looked at propagating sparse image labels to build predicted dense labels using cheaply acquired sparse image labels [77].

To acquire the graph information obtained by the model, most researchers have created models predicated on the premise that labels and characteristics change gradually throughout the edges of a graph. On the label side, node labels are propagated and aggregated via edges in the graph, which is known as the Label Propagation Algorithm (LPA) [82, 83, 84]; on the node side V, node features are propagated and modified through neural network layers, which is known as Graph Convolutional Neural Networks (GCNNs) [85, 86].

The labels of tagged data may spread across the edges, labeling all of the nodes in the network. Hence, a number of LP algorithms and graph regularization techniques have been proposed for various semi-supervised problems. Recently, Iscen *et al.* [87] revisited the LP method for semi-supervised learning, using an iterative approach of pseudo-labeling and network retraining to improve the performance of the algorithm.

2.3.2 Inexact Supervision

Imprecise supervision refers to when some supervisory information is provided, but it is not as precise as required. This occurs when only coarse-grained label information is provided. The assumption is that the training set is made up of labeled *bags*, each of which is made up of unlabeled examples. A bag is labeled as favorably if at least one instance contained inside is positive, on the other hand, it is labeled unfavorably if all the cases included within are negative. The purpose is to make educated guesses about the contents of hidden bags. This is known as Multi-Instance Learning (MIL) [88].

Dietterich *et al.* [89] first introduced MIL as a solution to the issue of pharmacological activity prediction. Unlike traditional supervised learning, MIL can deal with problems that have just partial label information, such as the label on a bag of examples. A number of novel methods [90, 91] have recently been proposed to address MIL issues by mapping each bag to a single point in a new feature space. These techniques convert MIL issues into a conventional learning problem that can be solved using single-instance classifiers.

2.3.3 Inaccurate Supervision

Inaccurate labeling is often the result of aggregating public or crowdsourced data sets [92]. The goal of weakly-supervised methods with noisy or inaccurate labels is to discover any occurrences that may have been mislabeled and to rectify or delete them. Numerous studies on inaccurate supervision have been suggested by the research community.

Noisy label learning has mostly been tackled using supervised learning [93]; nevertheless, how to cope with little labeled data and huge quantities of unlabeled data has yet to be thoroughly investigated. In the same vein as noisy label learning, instance-independent noise is studied initially [94]. In the broad case of convex surrogates, these main studies offer assurances for risk reduction under random classification noise. Instance-dependent noise [95] is a lot closer to the actual world, where label noise is determined by the inherent character of instances. Previously proposed methods have addressed the label noise issue by using inversed noise rates to create significant re-weighting algorithms for classification with label noise [96].

Scott *et al.* [97] investigate the classification issue in the context of a class-conditional noise model. They take an alternative approach to the issue, assuming

that the noise rates are unknown and that the current distribution fulfills a particular *mutual irreducibility* condition. They also don't provide an efficient solution to the issue.

Noise Regularization

It is difficult to train neural networks using just noisy labels since deep neural networks, in general, have a great capacity for fitting and memorizing the noise. The potential of enhancing a neural network's generalization capacity by adding additive noise to the training samples is addressed by several studies [98]. On the other hand, mathematical statistics have also been utilized to establish different asymptotic consistency findings. The study proposes mathematically justified criteria for selecting noise characteristics.

Because training samples change all the time, adding noise guides the network to learn better-discriminating patterns rather than memorize them from the input data [99], resulting in smaller network weights and a more resilient network with reduced generalization error. The noise makes it seem as though fresh samples are being pulled from the domain in the area of existing samples, smoothing the input space's structure. This smoothing may make it simpler for the network to learn the mapping function, resulting in better and quicker learning.

CHAPTER 3

OVERVIEW OF THE FRAMEWORK

3.1 Framework Overview

New technological advancements have led to massive amounts of data collected, which has profoundly affected many research and engineering fields. Numerous research possibilities need the creation and development of sophisticated, extensive data analytics methods. Given these needs, data science has emerged as a popular subject in both industry and academia, with applications ranging from real business solutions, technological breakthroughs, and interdisciplinary research to political choices, urban planning, and policy formulation. However, current tools and methods are far from adequate for gathering, organizing, and analyzing large amounts of data to find patterns, knowledge, and insight in using deep learning.

This doctoral dissertation tackles data science research problems by implementing an intelligent data analytics platform based on deep learning. The suggested study tackles many possible difficulties in a data science project, such as the availability and quality of hand-labeled data and developing a viable technique for modeling significant connections between features derived from different data sources. Figure 3.1 depicts the dissertation's overall framework, which comprises four primary components: multi-source multi-modality information fusion, fine-level pattern modeling with deep feature fusion, weakly-supervised training, and 3D advanced visualization. The overall framework has also been utilized in various applications, such as fluid dynamics, geometric morphometrics, building damage classification from satellite pictures, disaster scene description, and storm-surge visualization.

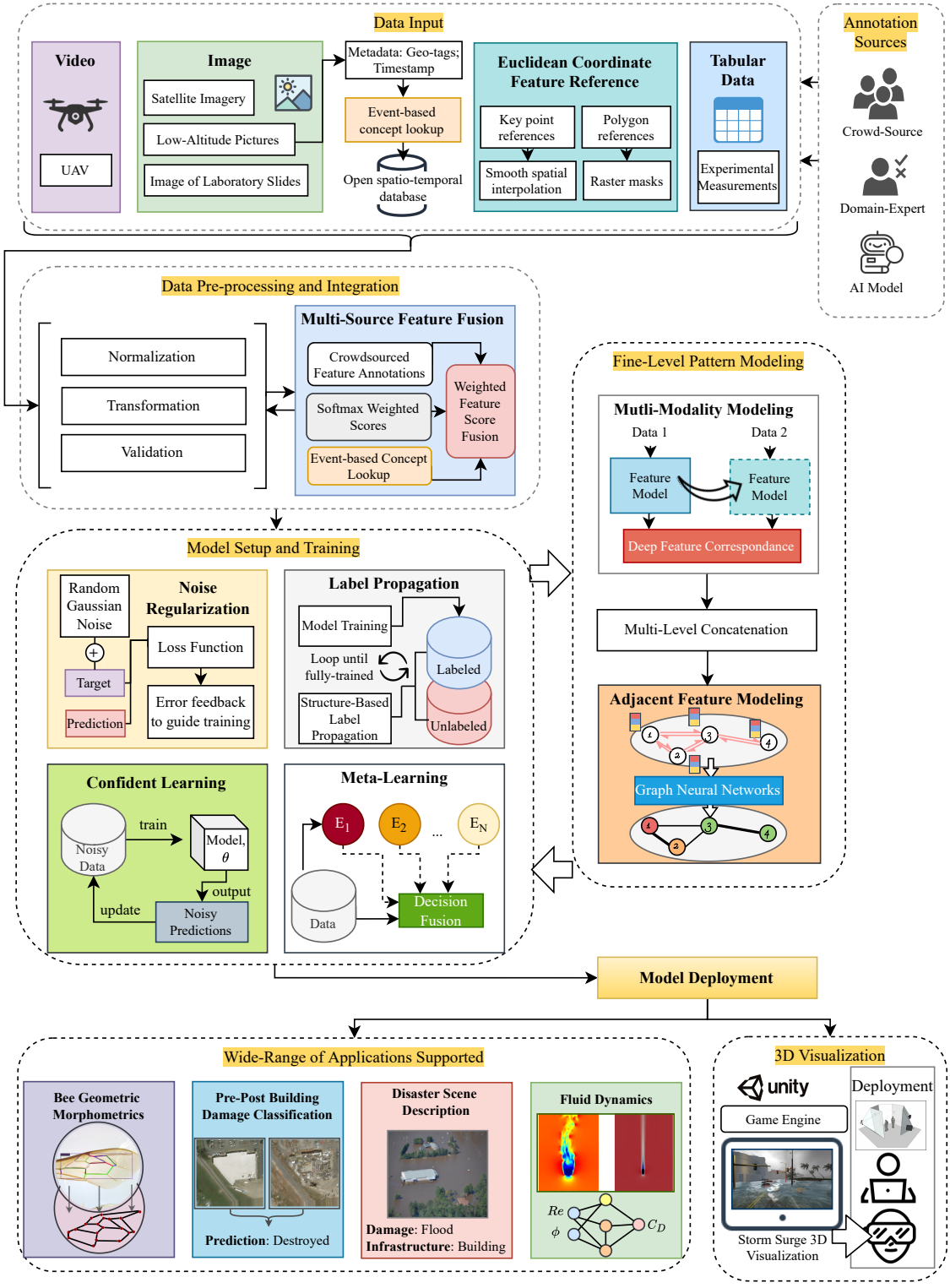


Figure 3.1: Overview of the dissertation's framework.

3.1.1 Multi-Source Multi-Modality Information Fusion

Multi-modal data has emerged as a significant area of big data, with each modality encoding a unique characteristic of data objects [100]. Different modalities are often complementary, prompting a lot of research into integrating multi-modal feature spaces to characterize these data objects better. Data fusion combines information from various modalities and sources to increase prediction accuracy. The significant data variations observed in these multi-source and multi-modal datasets, on the other hand, offer substantial technical challenges for machine learning and deep learning algorithms. We developed the intelligent data analytics framework with the capability to adapt to different data sources and modalities and make the proper transformations to prepare the data for fusion and modeling. This component contains two main parts:

- We train a unique DNN architecture [101] to model particle-based energy systems combining the scientific knowledge of previously proposed empirical correlations and an expanded feature set that captures essential aspects of the single-particle experiment system.
- We develop a multi-source weak supervision fusion [102] strategy that is trained on a significantly imbalanced dataset annotated with noisy labels.

Our framework harnesses the variety and valuable knowledge provided by supplemental data from other sources that can describe the contextual information found in the input data. The gathered data may come in various forms, including images, tags, or concepts detected by classifiers pre-trained on more general data benchmarks. The compatibility between the data retrieved from different sources and input data is first established at the semantic level or by utilizing an apriori association that links the distinct entities. The discrete variables acquired will be first

represented as continuous vectors using the embedding method. Semantic similarity is applied first to measure the distance of these vectors based on the similarity of their meaning or semantic content. After determining the semantic similarity, the data is combined using the proper aggregation and normalization techniques to fuse the information for training. Other compatibility methods to match the data are also proposed based on the information about the natural structure in real life. For example, spatio-temporal data can help match entities among compatible events and locations.

3.1.2 Fine-Level Pattern Modeling With Deep Feature Fusion

Fine-level pattern modeling aims to separate more nuanced characteristics amongst superficially similar groupings. When there are significant areas in the data from two or more highly similar classes, class overlap, also known as inter-class similarity, arises. Differentiating between two or more classes is difficult, if not impossible, due to considerable inter-class similarity. Other problems, such as class disparity, are worsened by the overlapping characteristics of the minority and majority classes. It's critical to extract fine-grained patterns from data to discover additional features that may help the model differentiate between different classes. This component contains three main parts:

- We extract local characteristics that automatically lead our proposed fine-level pattern modeling method to identify the honey bee subspecies by considering domain expertise data.
- We propose a two-stream CNN architecture [9] for recognizing buildings at four damage levels. We tested it using a curated, fully labeled dataset collected

from open sources. Our goal is to integrate the features generated by the two streams so that the overall model can comprehend how the inputs from both networks interact.

- We identify and fuse fine-level patterns from the data at multiple levels of the deep learning architecture, addressing both local and global interactions of the data's characteristics to produce a more accurate prediction for the samples with a high level of inter-class similarity.

A fine-level pattern modeling approach using deep feature fusion is proposed to provide a robust representation capability to identify such easily confused inter-class samples. A comprehensive approach is proposed for converting the additional information acquired through expert knowledge or metadata into the desired format that may be utilized to detect, filter, and magnify those fine-level patterns within the input data. The goal here is to integrate the localized features extracted from earlier layers to provide a more accurate prediction, especially for overlapping classes. We further use deep learning's capability to preserve the dominant structure of the input data, especially in the case of CNN, to extract the fine-level patterns from different layers at the same region of interest. Depending on the project goal, these local features may subsequently be concatenated with global features produced by later layers of the network that include more broad information about the input data.

3.1.3 Weakly-Supervised Training

In the case of supervised learning, annotated data is sent to a large number of recently established AI systems. When working with bigger models, however, identifying all of the data becomes more difficult. Furthermore, for a few professions that are just now starting to explore AI systems as a means of extracting valuable

information from raw data and aiding decision-making, there is simply insufficient tagged data. For these reasons, academics and practitioners are increasingly relying on less formal forms of supervision, such as heuristically generating training data from external knowledge bases, patterns, or other classifiers. This component contains two main parts:

- We develop using weak supervision in detecting and classifying damaged buildings [15]. The proposed research is distinctive in that it considers a scenario in which the only data available for training the model is an approximate estimate of the damaged building's location and degree of damage.
- This dissertation proposes a weakly-supervised deep learning system [19] that can handle noisy, restricted, and incorrect inputs while still predicting descriptive characteristics related to damage and the recorded environment. Because low-altitude imaging datasets like LADI are partially labeled, the soft-labels describing the probability of an image having a particular feature are propagated to unlabeled data in the training set to improve the training process.

We propose an end-to-end weakly supervised model where the assumption is that the label set is limited and might include instances of mislabels and errors. During training, random additive zero-centered Gaussian noise is introduced to create synthetic perturbations in the target label, augment the data, and minimize over-fitting. These minor perturbations are intended to assist in training a model that is resistant to the noise present in real-world data, such as coordinating position inaccuracies and crowd-sourced mislabels or biases. Our proposed framework combines a weakly supervised deep learning method with a novel label propagation model. It is possible to augment the training data using the label propagation method by adding labels to previously unlabeled data in order to increase the model's contextual awareness.

3.1.4 3D Advanced Visualization

We also demonstrate advanced visualization techniques that are used in a 3D game engine to provide an immersive and interactive experience based on multi-source data. This component, in particular, comprises the following:

- We look at a storm surge simulation [16] that was created utilizing a combination of real-world Geographic Information System (GIS) data and projected damages to create an automated virtual environment. Specifically, we use a 3D game engine to showcase sophisticated visualization methods to provide an immersive and interactive experience.

3.2 Dataset

3.2.1 xBD: A Large-scale Dataset of Satellite Imagery

Table 3.1: xBD dataset split is defined as the train, validation, and test split in this proposed work, along with each split's labeled polygon counts and proportion details for each label.

No.	Split	xBD Split	Polygon Counts	Label proportions (post-disaster)				
				no-dmg	minor-dmg	major-dmg	destroyed	un-classified
1	train	tier1 + tier3	316,114	0.739	0.081	0.068	0.074	0.037
2	val	test	54,862	0.755	0.087	0.070	0.069	0.018
3	test	hold	54,392	0.698	0.117	0.085	0.078	0.023

xBD [17] is used as a benchmark to assess the performance of our methods; it is a recently introduced large-scale dataset created for the advancement of building identification and damage assessment across multiple damage levels and damage types. xBD offers pre- and post-event RGB satellite imagery with building polygons, classification labels for damage forms, ordinal damage level labels, and corresponding

satellite metadata from a number of disaster events. The dataset contains around 700,000 building annotations across over $5,000$ $\texttt{km}^2$ of imagery from 15 countries.

The details of the xBD data are summarized in Table 3.1. We utilized the data from tier1 and tier3, which make up 80% of the overall data, to train the model. The test split was used as the validation set and the hold-out set as an unbiased test of the final trained model. DIU made tier1, and tier3 data split at the beginning of the competitions. The test split, despite its name, served as a validation set where the models that achieved the highest performance were selected. Competitors did not have access to the ground-truth annotations of the xBD's test split; however, results could be uploaded to an online leader-board system to measure a model's performance and improve accordingly. At the end of the competition, the best-performing models are further compared by utilizing the holdout set. Hence, the proposed study followed this split scheme.

Due to the competition's chosen pixel-level metrics, semantic segmentation has been the preferred approach for previously proposed building damage assessment research aiming to achieve the best results. Semantic segmentation assigns a class label to each pixel in an image. Multiple objects of the same class are treated as a single entity. However, this approach may run into the disadvantage of not being able to answer the questions in regards to how many individual buildings sustained a certain level of damage.

Data Challenges

Some characteristics, such as the type of buildings, are not included in the xBD dataset. The proportion of buildings in each class is grossly imbalanced, with the bulk of structures in the no damage class in most catastrophes. Because of the satellite's orientation changes, the building outlines identified on pre-disaster im-

ages may be altered. Images captured before and after a disaster event will look quite different due to variations in satellite characteristics (resolutions, angles, and altitudes).

3.2.2 Irma: Aerial Photographs from the Florida Keys

Figure 3.2: Samples of aerial photographs depicting the different levels of damage caused by Hurricane Irma on the Florida Keys.

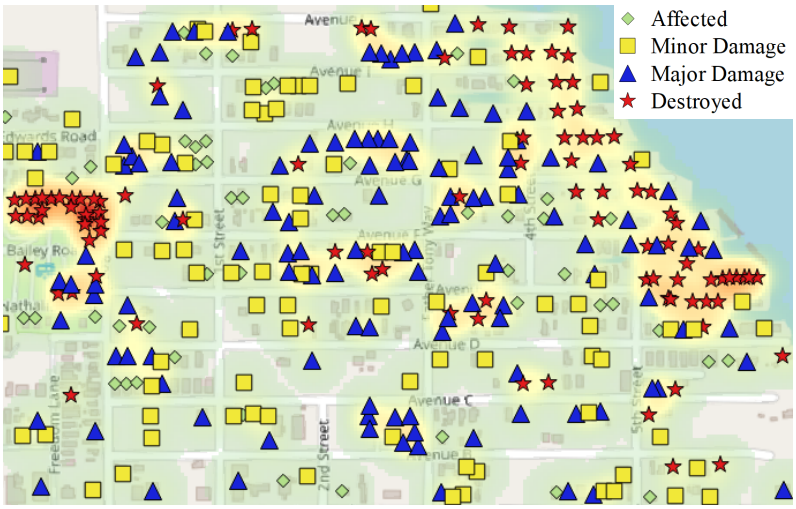


Figure 3.3: Irma damaged building location and damage level visualization from the Big Pine Key south area, one of the most affected regions after Hurricane Irma.

The Irma dataset demonstrated in Figure 3.2 of aerial photographs represents one disaster event focusing on wind damage. This dataset was processed and curated by our team and is composed of aerial imagery and damage assessment labels from open sources concentrating on the damages caused by Hurricane Irma in 2017 in the Florida Keys [103]. The aerial imagery was collected in the affected areas identified by FEMA and the National Weather Service during Hurricane Irma. The Monroe county area located in South Florida received most of the substantial damages [104]. Hurricane Irma struck the Florida Keys as a category four storm with a maximum sustained wind speed of 132 `mph` and storm surge reaching up to 8 `feet` [103]. The eye of the storm made landfall over Cudjoe Key, and consequently, the Lower and Middle Keys received the highest impact.

The Irma data comprises high-resolution aerial photographs representing a Ground Sampling Distance (GSD) of 50 *cm*. Alongside the aerial photographs taken before and after the disaster event, the damage assessment labels were also gathered by our team from public sources and combined and curated. Various damage assessment reports were combined from different data sources, including Monroe County's official public preliminary damage assessment report [105], Xian *et al.*'s assessment of the damages of more than 1600 residential buildings [106], and FEMA's Historical Damage Assessment Database [107].

The different damage levels are summarized as follows per FEMA's official guide to damage assessment [108]:

- *No damage*: The structure remains unchanged as seen from the birds' eye-view.
- *Affected*: The structure exhibits minimal effects, such as some missing shingles from the building rooftop, but it is still habitable according to FEMA's standards.

- *Minor damage*: It constitutes damages that do not necessarily affect the structure's integrity but may make it inhabitable until repairs are done.
- *Major damage*: The structure sustained substantial damage and required extensive repairs to make it habitable.
- *Destroyed*: The structure is a total loss where the repair will not be feasible for recovery.

Data Challenges

Unlike xBD, the Irma data includes the affected damage level, minor damage, major damage, and destroyed. As shown in Figure 3.2, damage labels from the Irma data are only the estimates of the damaged building's location and do not include the geometric information of the building footprint. The damage labels can be matched with building footprint geometries gathered from other data sources. Nonetheless, there is also the challenge of matching the location of the points to the suitable footprint.

3.2.3 Low Altitude Disaster Imagery

The Low Altitude Disaster Imagery (LADI) dataset [18] mostly consists of pictures captured from a low-flying aircraft by CAP and hosted by FEMA. The National Institute of Standards and Technology (NIST)'s TREC Video Retrieval Evaluation (TRECVID) competition released the dataset to participants in the middle of the year 2020. The LADI dataset uses a hierarchical labeling approach featuring five general categories, including *damage*, *environment*, *infrastructure*, *water*, and *vehicle*. Within each category, features of more specific categories are annotated. Different from the train set, which is mostly composed of still images obtained from

Figure 3.4: A visualization of the polygons and an indicated timestamp of when the picture was captured following the sequence.

an airplane, the LADI test set is a collection of short video clips captured from Unmanned Aerial Vehicles (UAV).

To include more concepts relevant to real-life events, we further utilize time and location metadata obtained from each image. Focal length (F), altitude (A), latitude, longitude, camera model, and so on are included in the information that can be extracted from the image metadata, assuming it follows the Exchangeable image file format (Exif). This information is useful to approximate the geographical region covered by the image taken from an airplane. Although there is no direct access via Exif to the height `H` and width `W` of the camera sensor, the camera model provided by the metadata was used to acquire this information from other sources. Simple trigonometry can specify the current footprint through the computation of width $= (A \times \texttt{W})/F$ and height $= (A \times \texttt{H})/F$ of the geographic region. The measured area is only a rough approximation of the area photographed, as illustrated in Figure 3.4, and is limited by its assumption that the camera takes the picture

while being pointed directly downwards. The angle from which the image was taken is a required parameter to be able to determine the exact geographical boundaries covered by the image.

Data Challenges

With only about 6% to 7% of the 500k images in the LADI training set being labeled by human workers, we tackle several challenges that arise from working with a highly-imbalanced and noisy dataset to train a reliable model. Moreover, some objects and features in the image are shown at different sizes and angles depending on the altitude of the picture, making some of these features difficult to detect.

3.2.4 Morphometric Bee Data Banks

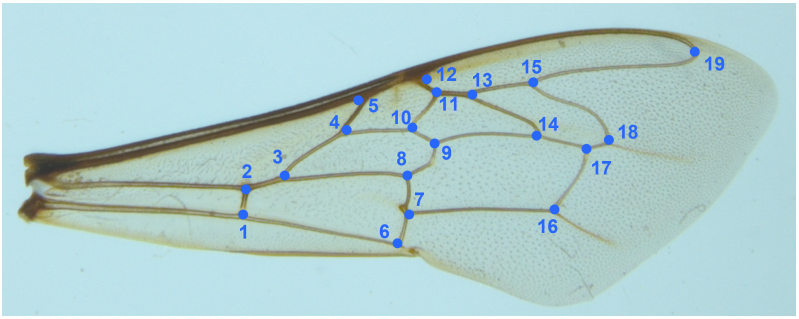


Figure 3.5: Forewing points of reference for geometric morphometric analysis.

Morphometric methods attempt to measure size, form, and size-form relation (allometry). The quality of landmarks in terms of comparability is of significant significance for comparisons of form. In concept and statistics, two distinct methods are (1) landmark morphometrics, which uses the relative location of a few anatomical sights, and (2) outline morphometrics, which capture shapes by use of a series of narrow pseudo-marks. Digitizing and integrating geomorphometric, genetic, behav-

ioral, and environmental data for individual insect specimens allows researchers to access data from various geographic locations, determine population changes, refine species identification tools, speed up the detection of evolutionary changes, and increase data accessibility. The model was validated using the curated Morphometric Bee Data Banks from various sources. Geometric morphometrics is a set of techniques for describing biological forms mathematically using geometric descriptions of their size and shape [109]. For honey bees, these morphological features are nodes at the junction of wing veins that relate to landmarks [12]. These wing morphological characteristics have been effectively employed to distinguish honey bee species and subspecies. In some instances, the geometric morphometrics approach may be as precise as more complex mitochondrial techniques [110].

3.2.5 Particle Drag Data

We compiled a dataset of 4202 data points from 29 previous publications, summarized in Figure 3.6. The selected datasets provide a complete set of features, allowing us to conduct unbiased tests of the trained models by isolating separate studies. At the same time, the rest of the data was used to train the model. Both regular and irregular particles are included in this dataset. Regular particles are particles with specific shapes that can be precisely described, such as cubes or cylinders. Irregular particles are natural particles with arbitrary shapes, such as grains or sands. The data was gathered from experiments that took place as far back as 1928 [126]. Because so many new definitions have been developed over the past 100 years, it was necessary to recalculate Re and C_D based on the terminal velocity stated in more recent publications in order to ensure that all of the data was held to the same standards. The drag force is then defined as a single function of the terminal

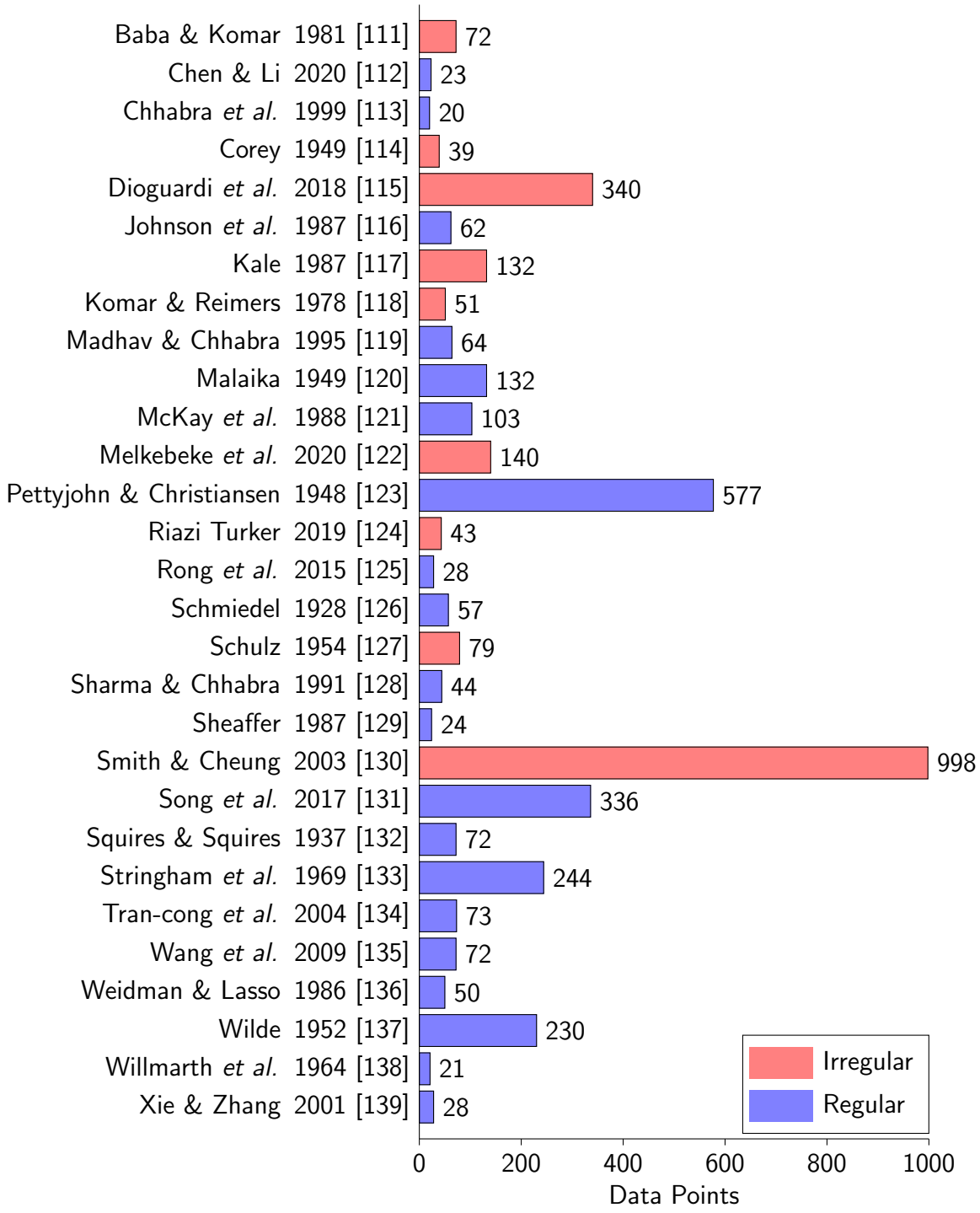


Figure 3.6: Summary of the collected experimental particle drag data colored by the particle shape.

velocity as shown in Equation 3.1.

$$C_D = \frac{4(\rho_p - \rho_f)d_p g}{3\rho_f u_t^2} \tag{3.1}$$

Based on the provided parameters, we attempted to compute the cross- and length-sphericity and particle sphericity with the most significant accuracy possible. Irregular particles were assumed to be ellipsoid when calculating their surface area. As a result, irregular particle sphericity is only an estimate of its true sphericity. The majority of these data were collected from single-particle free-fall physical experiments. In such an experiment, a particle is dropped into a tube filled with static Newtonian fluid. The drag force keeps increasing with the velocity (u_t) of the particle and eventually reaches a terminal velocity when the gravity of the particle balances the buoyancy force and the drag force.

CHAPTER 4

MULTI-SOURCE MULTI-MODALITY INFORMATION FUSION

Data from different sources and modalities contain a lot of valuable information [140, 141, 142, 143]. We aim to fuse this information to enhance the training process and extract more knowledge from the unstructured data by leveraging the metadata information. This metadata can be valuable to retrieve further information related to the input data from open databases about specific groups or equivalent entities related to the input data. Nonetheless, the data gathered from different sources often do not follow a standard structure that is easy to process. Existing deep learning approaches can be employed effectively to extract discriminating characteristics from most data because of its inherent structure, such as the grid-like organization of images.

On the other hand, further pre-processing must be applied to prepare the supplemental data for analysis and proper fusion with the input data. Furthermore, information learned by an AI model from more generic data standards may aid in knowledge transfer. Such supplemental data is not easily accessible or is very limited for a particular application. Depending on the research goal and the availability of the data, an intelligent approach must adapt to different data characteristics and challenges. This chapter describes many solutions to harnessing the knowledge from different data modalities and sources through fusion at the data level. Specifically, the proposed methods are evaluated in two challenging applications. The first application is a DNN trained to model particle-based energy systems combining the scientific knowledge of previously proposed empirical correlations. A multi-source weak supervision fusion strategy is used in the second application, which was trained on a significantly imbalanced dataset annotated with noisy labels.

4.1 Scientific Knowledge Aided Deep Neural Network Model

We propose the Drag Coefficient Correlation-aided Deep Neural Network (DCC-DNN) architecture to predict the particle drag force coefficient from various single-particle experimental data. Beyond sphericity and Reynolds number, the proposed approach includes an expanded set of features supported by the literature. These features are density ratio, aspect ratio, and lengthwise and crosswise sphericities. Simultaneously, model regularization and meta-learning help train a generalized and more reliable drag model, despite the limited data available and the variance exhibited in individual single-particle studies. A comprehensive ablation study demonstrates the suggested method's advantages over alternative state-of-the-art solutions. The presented model applies to spherical and non-spherical particles, providing much-needed generality and reliability for industrial application.

Particles are encountered in various engineering and natural processes, including multiphase energy systems, an integral component of fluidized beds. Fluidized bed technologies have outstanding heat transfer characteristics, enabling them to produce syngas from various solid fuels toward lowering carbon pollution and promoting environmental sustainability. Other processes that involve particles include sand ingestion in air-breathing engines, volcanic debris ballistic flight in the atmosphere, and sedimentation. The common presence of many particles in these processes makes such flow problems complex and computationally expensive to model, often necessitating techniques that relax certain physics [144, 145].

Modeling such particle-based energy systems requires the knowledge of the drag force coefficient (C_D), a non-dimensional number that relates the actual drag force on the particle to the reference quantities such as relative velocity and fluid density, and cross-sectional area. Although researchers in this field have reported numerous

data on C_D, the relationship between C_D and the particle and fluid features remains uncertain, particularly for non-spherical particles at a higher Reynolds number (Re).

A particle's C_D is a dimensionless coefficient that is used to measure the drag impact on a particle in a fluid-particle system $F_d = 1/2 C_D \rho_f u_t^2 A$, where F_d is the fluid force (or drag force) exerted on a particle, ρ is the density of the fluid, u_t is the velocity of the particle, and A is the reference area of the particle. However, deriving a numerical model for a particle's drag force is notoriously challenging. The only such solution is the renowned Stokes' law $F_d = 3\pi d\mu u_t$, where d is the diameter of a sphere and μ is the dynamic viscosity of the fluid. Combining the two equations, one can quickly obtain a solution for the drag coefficient, $C_D = 24/Re$. This equation, however, is only applicable for a single spherical particle and only when Re approaches zero. As a result, considering a non-spherical particle or a larger Re scenario, the drag coefficient/force must typically be determined experimentally. For moderate to large Re, the particle shape substantially influences the drag coefficient value. The form is often quantified by sphericity (ϕ), defined as the ratio of the volume equivalent sphere's surface area to the particle's total surface area.

Early studies have reported correlations to account for the shape factor using experimental and numerical approaches [146, 147, 148]. However, these studies are limited to certain types of particle shapes and Re ranges. Hence, a more sophisticated technique is required, such as a neural network that can be generalized to a wide range of single-particle experiments. Recently, academics have developed an interest in Machine Learning (ML) and Deep Learning (DL) approaches for predicting C_D [149] using either experimental or simulated data or a combination of the two. However, several challenges to estimating C_D are still several challenges, including inadequate data available within a wide range of features and data vari-

ability observed from different experiments. Due to the difficulties of generalizing to new data points, few publications have attempted to address the challenge of training deep architectures with a limited number of samples.

4.1.1 Motivation

There is no analytical solution to the particle drag force coefficient other than the famous Stokes' Law, where extremely low Re and spherical particle assumptions are made. Experiments mainly achieve the single-particle drag force coefficient of non-spherical particles at larger Re. Relationships between the drag force coefficient and particle and fluid parameters are unknown. Machine Learning (ML) techniques are widely believed as an effective strategy for developing predictions when the relationships between input and output variables are unknown. Early studies, to name a few [150, 151, 152, 153], have proven that ML techniques are a viable way to predict key parameters in particle sediment studies.

With the recent advancement of deep learning techniques, more studies use deep neural networks to predict particle settling related parameters. He and Tafti [149] predicted sphere particle drag force coefficient using a single hidden layer Artificial Neural Network (ANN). The input variables were Re, volume fraction, and relative neighboring particle locations obtained from numerical simulations. The Levenberg-Marquardt algorithm trained the ANN. 52% of the prediction was within a relative error of 10%. Yan *et al.* [154] predicted the drag force coefficient of non-spherical particles using a Radial Basis Function (RBF) network and a Back Propagation Neural Network (BPNN). Only sphericity and Re obtained from previous publications were used as the input variables. The result was well fitted for spherical particles but much less accurate for particles with low sphericity. Chen and Li [112]

predicted particle drift velocity and drag force using a three-hidden-layer Deep Neural Network (DNN). Volume fraction, particle slip velocity, and pressure gradient were used as the input variables. Adam was used for training the DNN, and a good accuracy was attained. Balachandar [155] directly predicted the drag force of a spherical fluidized bed based on Re and volume fraction using linear regression. An R-squared (R^2) error of 0.64 was achieved. Zhu [156] predicted the drag correction factor of a fluidized bed based on volume fraction, particle size, pressure gradient, etc., using a three-hidden-layer DNN. Adam trained the DNN. A Mean Absolute Percentage Error (MAPE) of 9.85% was achieved with this network.

Hwang *et al.* [157] proposed a CNN framework to predict the drag, lift, and torque coefficients in the low Re regime based on the PR-DNS (Particle Resolved-Direct Numerical Simulation) simulation data. To avoid heavy computation while keeping an accurate particle geometrical information, the Variational Auto-Encoder (VAE) was utilized to extract the shape and orientation of a particle. VAE was then compressed and correlated to the force obtained from the PR-DNS data. Luo *et al.* [158] developed a feedforward neural network to predict the drag force in a fluidized bed based on Re, volume fraction, velocity, and position fluctuation of particles obtained from the DNS (Direct Numerical Simulation) simulation. The model was compared with other empirical correlations, and a simplified equation was achieved based on the trained network. Several ML models and the importance of various input features were also evaluated by Rushd *et al.* [159].

We developed a model that can predict C_D values using multiple features while also considering new and unknown data. The neural network can efficiently consider the effects of flow and particle features and predict the C_D with high accuracy [112, 154]. Different from the existing works that use specific data sets for drag modeling, this chapter shows the development of a general drag model coined

as the DCC-DNN to estimate the C_D using an expanded range of geometrical features. The findings indicate that utilizing a larger feature set, regularization, and a meta-learning method makes it feasible to predict C_D with satisfactory accuracy.

The new contributions of the present research are as follows:

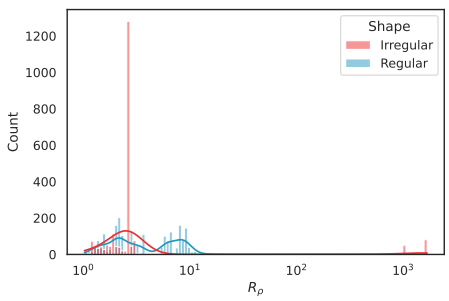
- This chapter introduces a novel DNN architecture for combining the scientific knowledge of previously proposed empirical correlations with neural networks to develop a general drag model.
- As far as the authors know, the dataset used to train and evaluate the proposed approach is the largest compared to previous studies. Moreover, the proposed technique incorporates an expanded range of features supported by the literature.
- Extensive ablation tests are done in this study to show how the proposed approach is superior at getting the best results and quickly adapting to new data.

To the best of our knowledge, developing a general drag model that can adapt to different experiments and scenarios within wide parameter ranges has yet to be explored and reported in the literature.

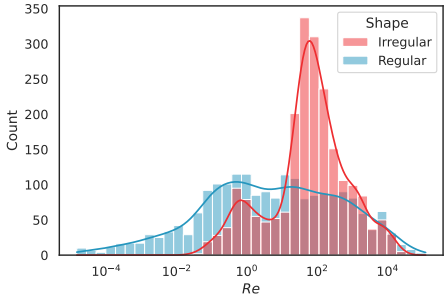
4.1.2 Proposed Methods

Feature Engineering

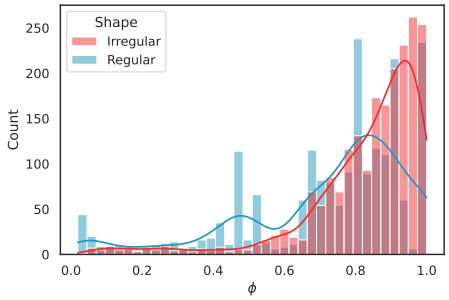
This section will discuss how the proposed expanded feature set captured essential aspects from the single-particle experimental data, including flow property, particle geometry, and settling direction. Along with the traditional Re, sphericity ϕ, density



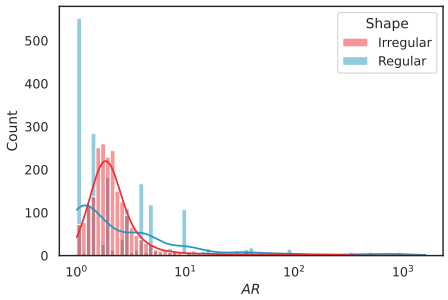
(a) Density ratio



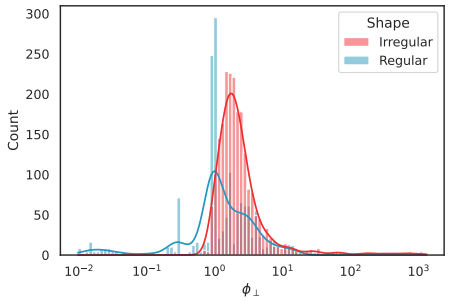
(b) Reynolds number



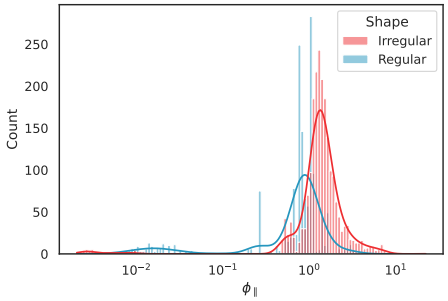
(c) Sphericity



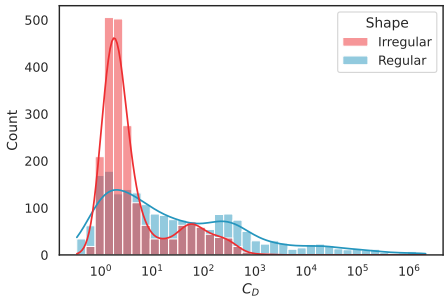
(d) Aspect ratio



(e) Crosswise sphericity



(f) Lengthwise sphericity



(g) Drag values

Figure 4.1: Feature distribution plots are categorized by the generic particle shape.

ratio R_ρ, aspect ratio AR, crosswise sphericity $\phi_\perp$, and lengthwise sphericity $\phi_\parallel$ are some of the features included in this study to improve the modeling approach.

Flow Property: Reynolds number Re and density ratio R_ρ are applied as input features to measure the flow property of the particle samples and predict the C_D. The Re is defined as $(\rho_f w_p d_p)/\mu_f$. The density ratio R_ρ is defined as the ratio between fluid density and particle density, i.e., ρ_f/ρ_p.

Particle Geometry: Two features are used in this study to define particle geometry. Both regular and irregular particles are described with three principal axes l_l, l_m, l_s (long, medium, and short axes). The aspect ratio (AR) of a particle is defined as the ratio of l_l over l_s of a particle, $AR = l_l/l_s$. As demonstrated in Equation 4.1, the ϕ is calculated as a fraction between the surface area of a volume of an equivalent sphere and the actual particle surface area. For regular particles, since both volume and surface area can be precisely measured, the sphericity can be easily achieved. For irregular particles, the measurement of the surface area is extremely difficult. Therefore, all irregular particles are approximated as ellipsoids, and the surface area (A_p) is calculated through Knud Thomsen's formula [160], as shown in Equation 4.2.

$$\phi = \frac{\pi^{1/3}(6V_p)^{2/3}}{A_p} \tag{4.1}$$

$$A_p = 4\pi \left(\frac{(\frac{l_l l_m}{4})^{1.6075} (\frac{l_l l_s}{4})^{1.6075} (\frac{l_m l_s}{4})^{1.6075}}{3} \right)^{1/1.607} \tag{4.2}$$

Settling Direction: As demonstrated in Equation 4.3, crosswise sphericity ($\phi_\perp$) is measured as the proportion of the volume equivalent sphere's cross-sectional area to the particle's projected area perpendicular to the flow. On the other hand, lengthwise sphericity ($\phi_\parallel$) is the proportion of the volume equivalent sphere's cross-

sectional area to the difference between half the particle surface area and the particle's average longitudinal projected area, as shown in Equation 4.4 [148].

$$\phi_{\perp} = \frac{\pi^{1/3}(6V_p)^{2/3}}{4CSA_{\perp}} \tag{4.3}$$

$$\phi_{\parallel} = \frac{\pi^{1/3}(6V_p)^{2/3}}{2A_p - 4CSA_{\parallel}} \tag{4.4}$$

The longitudinal direction is the particle's settling orientation. However, since some of the experimental data utilized in this research did not disclose the particle settling orientation in their experiments; alternative assumptions were made. Several representative shapes, including cuboid, cylinder, tetrahedron, prolate, and oblate ellipsoids with a wide range of AR were tested under a low Re regime. After that, the particle settling behavior was divided into three categories: across, lengthwise, and diagonal. Crosswise particles settle along the shortest principal axis, lengthwise particles along the longest main axis, and diagonal particles along the diagonal line. The particle cross-section area was calculated using the projected settling orientation. For a given inertia matrix, we simulate the particle spinning under the effect of a torque vector. The torque vector and inertia matrix are adjusted at each step to account for flow characteristics and direction changes. After a few steps, the particle's orientation stabilizes, and the particle's net torque converges to zero or oscillates around zero. This simplified approach enables a rapid prediction of the particle's settling orientation in a low Re area.

Deep Neural Networks

Neural networks are a machine learning approach inspired by the inner biological workings of the brain. It is no surprise that neural networks have gotten a lot of attention recently after delivering breakthrough results in image recognition, voice

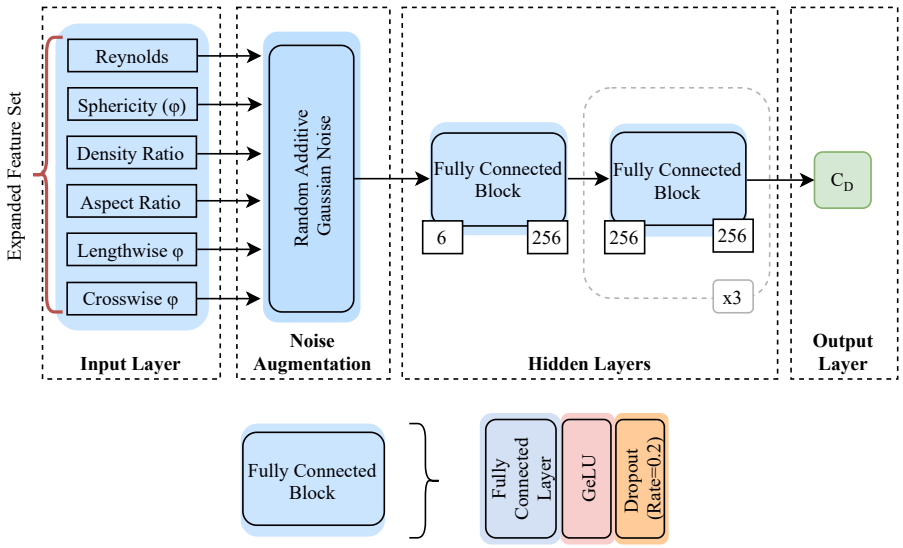


Figure 4.2: The proposed single-model DNN architecture. Each block's input and output sizes are indicated by the numbers in the squares.

recognition, and natural language processing [1]. Deep learning is well-known for using multiple-level representation learning–automatic feature identification from raw input data. Each layer of the network alters the data, generating a new representation and allowing the data to be separated linearly. Figure 4.2 illustrates our novel DNN model that takes as input an expanded set of features and predicts C_D. During training, a noise augmentation layer adds random Gaussian noise to the input feature set to assist the model in dealing with the errors that may be found in the data. The hidden layers are grouped into four fully-connected blocks composed of a fully-connected layer, an activation function, and a random dropout layer.

Activation Function: In our DNN model, the activation function instructs each neuron on transforming input before passing it on to the next neuron in the following processing layer. Our proposed DNN applies a relatively new and high-performance neural network activation function known as Gaussian Error Linear Unit (GELU). In contrast to the traditional Rectified Linear Activation Units (ReLUs), the GELU nonlinear activation function weights inputs by magnitude rather

than sign. GELU has been previously tested on different datasets and regularly outperforms the Exponential Linear Unit (ELU) and ReLU, making it a viable alternative to the prior nonlinearities [161].

Loss Function: The problem of training a model is solved by minimizing the loss function. Loss functions are required in every statistical model because they offer a benchmark against judging the model's performance. Moreover, the parameters learned by the model are established by minimizing a given loss function. To tackle the regression problem, we assess three loss functions that are resilient to noise and outliers: Mean Absolute Error (MAE), Log-Cosh, and Huber. While all three may be used to train a model to predict real-valued data by lowering the average loss between actual and anticipated values, they each perform better in different circumstances.

Model Regularization

Regularization is a collection of strategies and procedures designed to solve the problem of over-fitting by decreasing the model's generalization error while maintaining little impact on the training error. This paper shows two fundamental strategies to assist our model in learning the underlying patterns from the training data rather than memorizing random information.

When a neural network is trained on a small dataset, it may memorize all training instances rather than learning the underlying pattern, resulting in overfitting and poor performance. Due to the irregular or sparse sampling of points in the high-dimensional input space, small datasets may also provide a more difficult mapping issue for neural networks to solve. One strategy for smoothing the input space and making it simpler to learn is introducing noise to the inputs during training. Zero-centered Gaussian noise [15] is randomly added to the input feature set while the

model is trained to introduce synthetic fluctuations into the values, augment the input data, and avoid over-fitting. These minor variations are designed to aid in developing a model that is resistant to the noise common in experimental data of this sort.

A dropout layer is applied for the output of each fully connected block, as shown in Figure 4.2. The idea for dropout [1] is to remove hidden network units stochastically during training. The inclusion of these two regularization techniques adds more variability and noise to the model, resulting in more generalized and reliable models.

Drag Coefficient Correlation-aided Deep Neural Network (DCC-DNN)

Traditional correlation-based drag models published in the literature are developed mainly for spherical particle data. For non-spherical particles, however, these correlations are often restricted to extremely small ranges of particle shape factors and flow conditions. This study explores two meta-learning techniques, Stack Generalization and Mixture of Experts, to incorporate the previously proposed scientific knowledge found in empirical equations to develop the proposed DCC-DNN. More specifically, four well-known equations [146, 147, 162, 148] have been selected to demonstrate how these empirical correlations can aid the model in obtaining better performance. These meta-learning techniques aim to dynamically aggregate predictions from many predictive models to achieve the highest accuracy.

Stack Generalization: Stacked Generalization (SG), or stacking, learns how to integrate the predictions from two or more machine learning algorithms. The advantage of stacking is that it may combine the skills of several high-performing models to generate predictions that outperform any single model in the ensemble on a classification or regression test. Studies such as Rushd *et al.* [159] have started to

explore the SG technique to combine different classification models and predict the terminal settling velocity of a particle. Stacking, on the other hand, can only learn constant weights based on an optimization criterion, which may be problematic for generalization.

Mixture of Experts: The Mixture-of-Experts (MoE) [163] technique entails creating a gating network that learns which expert's prediction to trust depending on the input features and then combining the predictions. For each data sample, the model is able to choose just a subset of experts that can best predict C_D via the gating network conditioned on the input feature set. Our gating network includes a softmax output that provides each expert with a probability-like confidence score. The final output is a weighted sum of all the experts' outputs. The gating network is configured similarly to the baseline model previously introduced.

4.1.3 Experimental Analysis

Dataset

We gathered a 4202-data-point dataset from about 30 prior studies as illustrated on Figure 4.3. The chosen datasets give a comprehensive collection of attributes, enabling us to isolate independent research and perform impartial evaluations of the trained models. The remainder of the data was utilized for training the model simultaneously. This dataset includes both regular and irregular particles. Regular particles, such as cubes or cylinders, have precisely specified forms. Irregular particles, such as grains or sands, are natural particles having random shapes. The data was gathered from experiments that took place as far back as 1928 [126]. Because so many new definitions have been developed over the past 100 years, it was necessary to recalculate Re and C_D based on the terminal velocity stated in more recent pub-

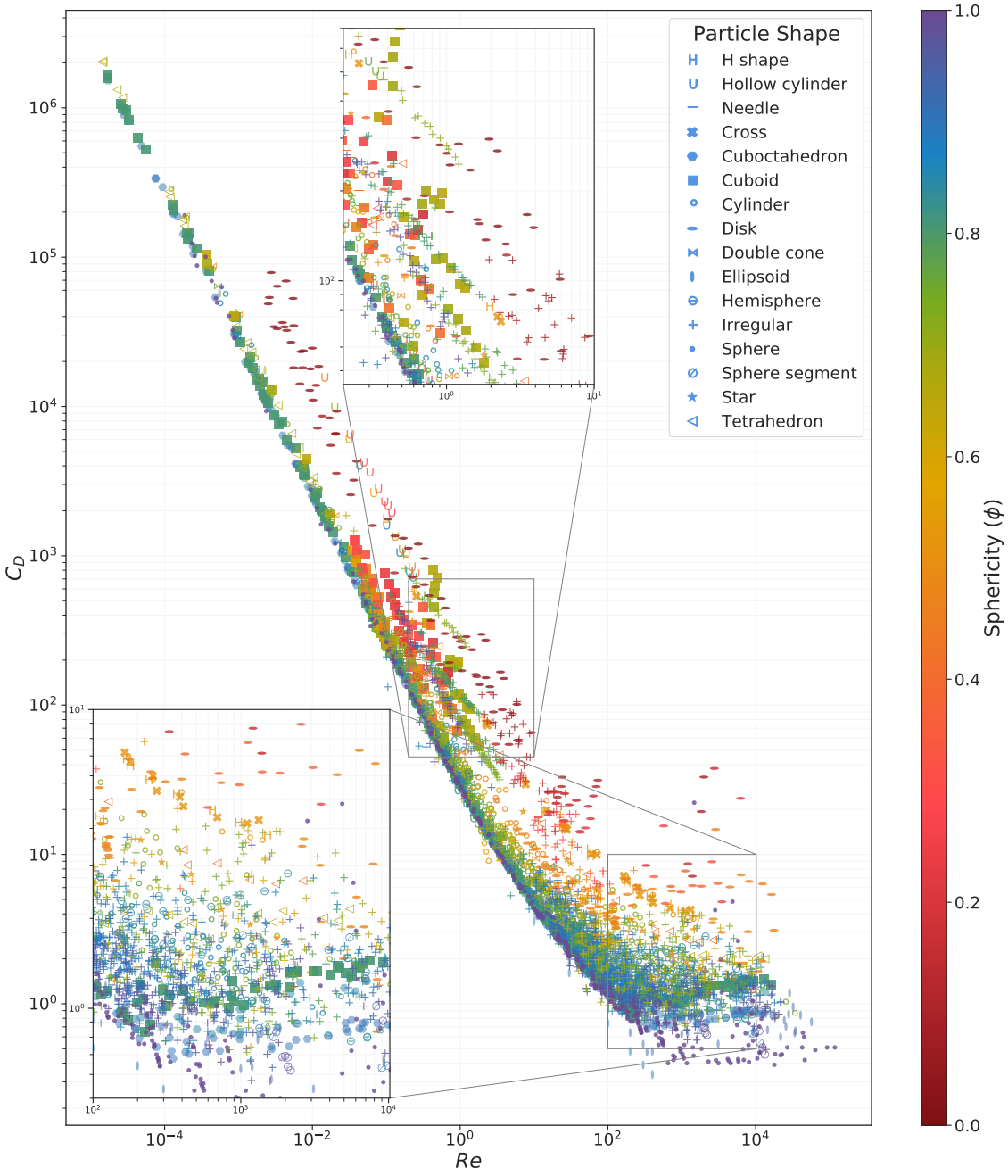


Figure 4.3: The effect of ϕ on the C_D as a function of the Re is illustrated using the data gathered from various studies in the literature.

lications in order to ensure that all of the data was held to the same standards. The crosswise and lengthwise sphericities and particle sphericity were calculated as ac-

curately as possible based on reported parameters. Irregular particles were assumed to be ellipsoid when calculating their surface area. As a result, irregular particle sphericity is only an estimate of its true sphericity. The majority of these data were collected from single-particle free-fall physical experiments. In such an experiment, a particle is dropped into a tube filled with static Newtonian fluid. The drag force keeps increasing with the velocity (u_t) of the particle and eventually reaches a terminal velocity when the gravity of the particle balances the buoyancy force and the drag force.

Experimental Setup

Data Preparation: Normalizing data by mean and standard deviation is most helpful when the data distribution is approximately symmetric. Because Re and C_D have skewed distributions, the natural log is employed to convert them such that their distributions are as near to a normal distribution as feasible before applying data pre-processing. We initially normalize features by removing the average and scaling to unit variance before training the model. Each feature is centered and scaled independently by computing the relevant statistics on the samples from the training set. Many machine learning and deep learning approaches require dataset standardization; otherwise, they will perform poorly if the individual features are not normally distributed [1]. The mean and standard deviations obtained from the training set are then recorded and used to alter future data.

We then conduct a k-fold cross-validation study to ensure the validity of the proposed model. For this type of validation, the training set is randomly divided into $k = 10$ subgroups containing an equivalent number of samples. The regression model is trained on $(k - 1)$ datasets before being tested on the remaining dataset. This technique ensures that each of the k subsets is used as a validation set at least once.

Our setup is different from the experimental setups found in the previous studies of the drag model. We construct folds that maintain the proportion of samples for the target values as feasible while adhering to the constraint of non-overlapping experimental sources between splits. This setup style allows us to validate the regression model and demonstrate the capability of the proposed technique to adapt to different experiments.

We deal with missing values via iterative imputation [164], a procedure in which each feature is represented as a function of the other features (using the Extra-Trees regression [165]), such as a regression problem where missing values are anticipated. Each feature is imputed sequentially, allowing for using previously imputed values as part of a model for predicting future features. It is iterative because this procedure is repeated many times, allowing for ever-improving estimates of missing values as all features' missing values are estimated.

Performance Evaluation: Many metrics are used to assess the model's performance statistically. These performance metrics were selected for the current study to allow a systematic comparison of the performance of different prediction models as well as the traditional empirical models. In our regression problem, the assumption is that y_i represents the target value for the i-th sample of the data, $\hat{y}_i$ signifies the predicted value, and N is the cumulative total number of sample points. Each of these performance metrics is defined as follows.

- **Root Mean Squared Error (RMSE)**: Residuals (i.e., predictive mistakes) are used to calculate the regression line's distance from the data points. RMSE is the standard deviation of these residuals, which shows the degree to which the data are concentrated around the line of best fit.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4.5}$$

- **Mean Relative Absolute Error (MRAE)**: Both extreme and low values are both susceptible to MRAE. The following formula yields the MRAE:

$$MRAE = \frac{1}{N}\sum_{i=1}^{N}(|y_i - \hat{y}_i|)/y_i \tag{4.6}$$

  The MRAE is very sensitive to outliers and low values.

- **Normalized Residual Sum of Squares (NRSS)**: Employed as a measure of variance within the residuals, a low NRSS suggests that the model is well-fit to the data.

$$NRSS = \sum_{i=1}^{N}(\frac{(y_i - \hat{y}_i)^2}{y^2}) \tag{4.7}$$

- **Sum of Squared Log Error (SSLE)**: This metric is robust to outliers while calculating the relative error between actual and predicted values.

$$SSLE = \sum_{i=1}^{N}((log(y_i) - log(\hat{y}_i))^2 \tag{4.8}$$

- **R-squared ($\mathbf{R^2}$)**: This is also referred to as the coefficient of determination to indicate how near the original data is to the fitted regression line.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4.9}$$

Results and Discussions

Single-Model Ablation Study: We first conducted an extensive ablation study on the single-model DNN (previously illustrated in Figure 4.2). This ablation study shows the best configuration for each relevant hyper-parameter and the magnitude of the influence the parameter value has on the model's performance. During the ablation research, five factors are investigated to see how they affect the performance of the single model based on the DNN architecture. The model depth (number of blocks) and width (number of neurons) are among the components investigated

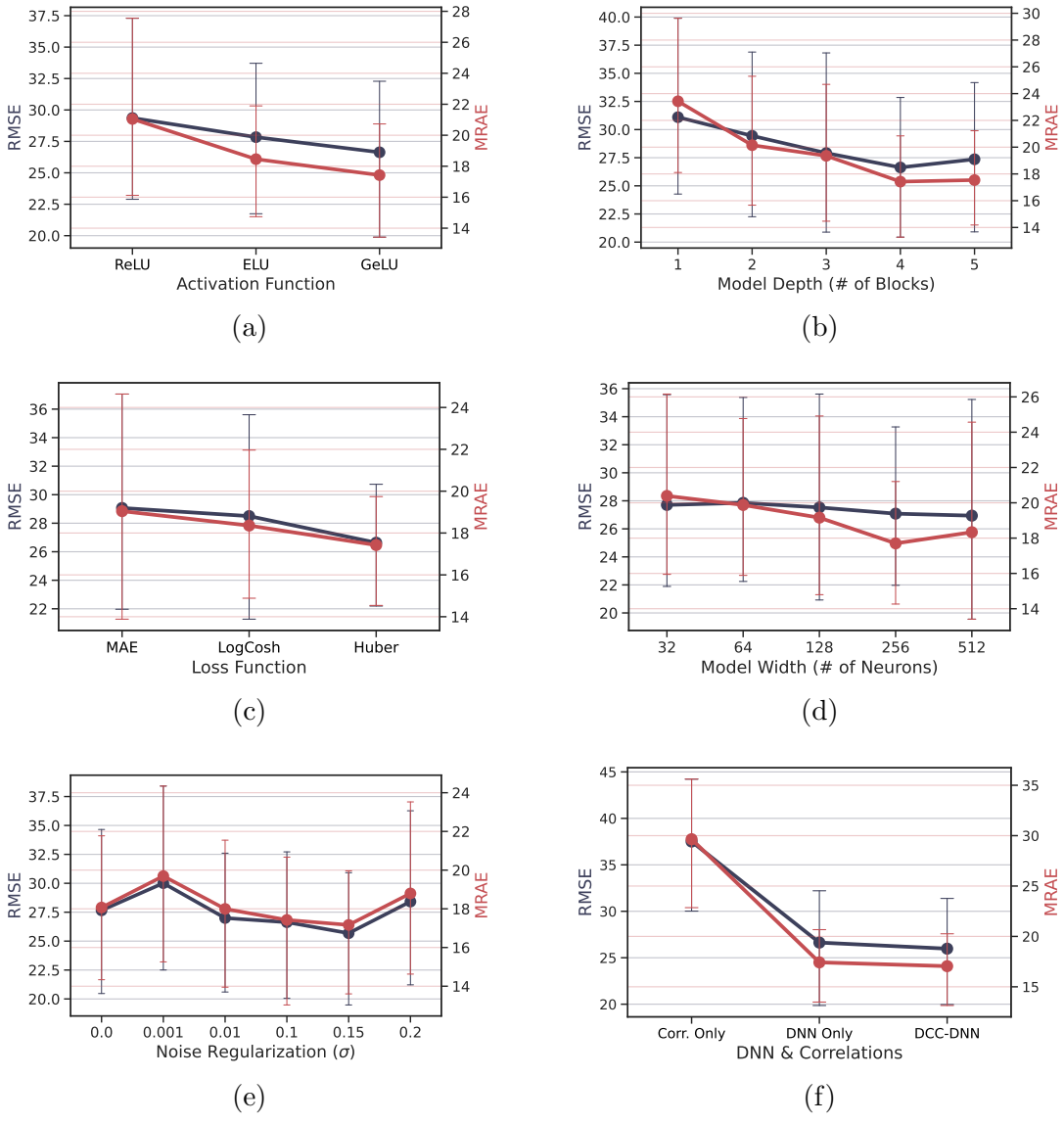


Figure 4.4: The results of the ablation investigation on the suggested single-model based on DNN are shown in the form of point plots. Each point is placed in the mean for the values calculated from two metrics (RMSE at the left in black and MRAE at the right in red) calculated from the cross-validation results. Mean error point values are plotted with the standard deviation bars at 95% confidence interval.

through the ablation study, along with activation function, loss function, and the degree of noise augmentation. Figure 4.4 illustrates the point plots for each of these components to estimate two error metrics' (i.e., RMSE and MRAE) central tendency based on the error value from each cross-validation fold and the details on uncertainty around that estimate through error bars with 95% confidence. The lines connecting each point from the same color level enable the interactions to be evaluated by variations in slope.

We can appreciate the significant impact made by each of the explored components to develop the final configuration for the proposed DNN model. Each of the chosen hyper-parameters for activation, loss function, and noise regularization has been studied and proven by the literature to help the model generalize better and handle noise in the data. Our model achieves a better performance applying the GELU activation, as demonstrated in Figure 4.4a, with four fully-connected blocks (Figure 4.4b), each containing 256 neurons, as shown in Figure 4.4d. Although it is very sensitive to outliers, the mean squared error (MSE) loss function is the most basic and widely used. Loss algorithms that are less susceptible to outliers, such as MAE, Log-Cosh, and the Huber loss, are compared to train a more resistant model to outliers observed in the input data, as shown in Figure 4.4c. By balancing the MSE and MAE together, the Huber Loss achieves the best performance by using MAE for bigger loss values and reducing outliers' weight. The importance of noise regularization in Figure 4.4e is further demonstrated to augment the training set and improve the model's capability of handling noise given $\sigma = 0.15$. Finally, using the MoE method, we compared the proposed approach's performance instead of only using the single-model DNN or the best-weighted combination from the empirical correlations.

Table 4.1: Performance measures of predicting C_D using cross-validation among traditional correlations (TC), well-known machine learning (ML) methods, and different deep learning (DL) configurations, including the proposed DCC-DNN, in which $S : < Re, \phi, \phi_\perp, \phi_\parallel, AR, R_\rho >$, represents the complete feature set.

Type	Method	Input Features	RMSE	MRAE	NRSS	SSLE	R^2
TC	Haider & Levenspiel, 1989 [146]	$< Re, \phi >$	37.93±12.13	30.08±11.08	56.06±54.48	19.52±21.10	0.7146±0.38
	Chien, 1994 [147]	$< Re, \phi >$	49.46±12.29	38.59±9.06	92.21±90.81	26.14±29.03	0.6259±0.40
	Yow *et al.*, 2005 [162]	$< Re, \phi >$	200.91±92.01	164.10±85.42	2001.56±2686.58	31.47±34.76	-1.5375±7.44
	Holzer & Sommerfield, 2008 [148]	$< Re, \phi, \phi_\perp, \phi_\parallel >$	55.13±28.29	46.26±24.90	111.61±127.72	46.39±51.91	0.1171±1.71
ML	Random Forest [165]	S	48.52±12.73	34.70±9.08	121.77±190.37	19.85±20.62	0.5426±0.33
	Gradient Boosting [165]	S	45.12±14.10	33.01±8.21	108.76±152.78	18.23±20.01	0.5891±0.31
DL	Baseline	$< Re, \phi >$	36.38±9.72	28.59±6.16	62.14±87.10	11.84±13.33	0.7971±0.22
		S	37.09±8.92	27.05±5.18	72.66±116.09	72.66±116.09	0.7822±0.20
	Single-Model DNN	$< Re, \phi >$	31.54±12.89	23.60±9.41	42.62±49.90	8.31±9.76	0.7508±0.30
		S	26.63±10.63	17.43±6.29	40.83±62.95	6.96±10.50	0.8118±0.25
	DCC-DNN (SG)	S	44.66±12.47	33.92±7.29	76.41±81.48	13.22±12.52	0.8150±0.19
	DCC-DNN (MoE)	S	**25.98±10.18**	**17.05±6.03**	**38.00±58.17**	**6.76±10.02**	**0.8569±0.20**

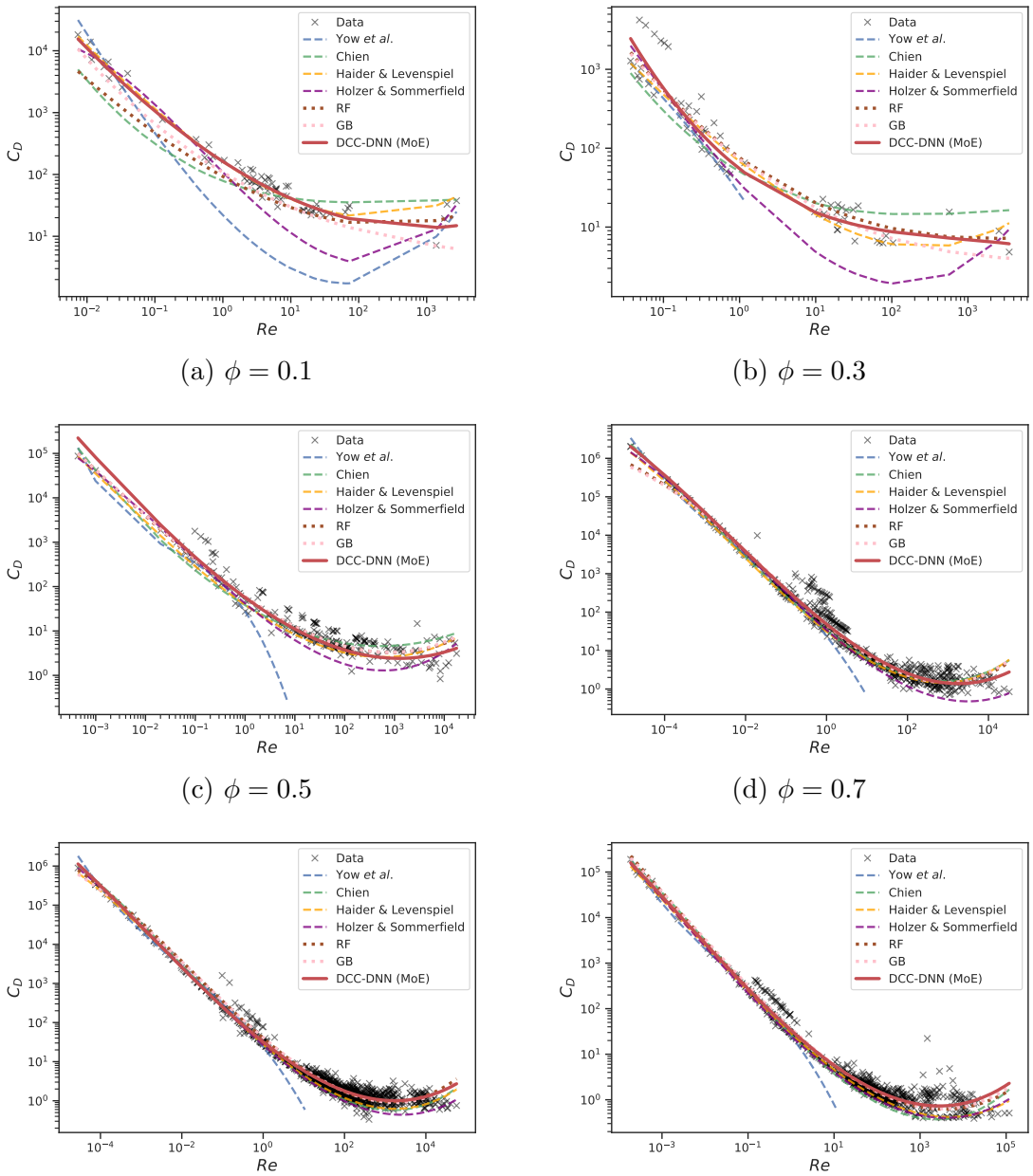
Proposed Model Comparison: Table 4.1 shows an overview of the performance of various model configurations using different predictive techniques (i.e., traditional correlations (TC), machine learning (ML), and deep learning (DL)). The average of the metrics previously introduced and acquired from the 10-fold cross-validation test sets is provided, followed by the standard deviation. The table also shows the input features used for each of the methods. Based on the numerical correlation from earlier research, we compare the results of our proposed method. Overall, we can appreciate that the DL models can predict better results when the complete feature set (S) has been used in training, showing the importance of the training data in predicting a more accurate C_D. The more data we can feed the model to learn, the better result we obtain.

Unlike ML and DL techniques, the TC methods are not fitted to the training set. The results for the TC methods are the summaries of their predictive performance on the cross-validation test set such that comparisons with other techniques are consistent. The other two ML models, Random Forest (RF) and Gradient Boosting (GB),

are tree-based techniques that have been extensively used in the literature and have demonstrated impressive results [165]. Nonetheless, their performance shows their inability to generalize and adapt to new data. The performance summary further shows different DL configurations starting with the baseline model (configured as one hidden layer of 64 neurons employing ReLU activation and optimized through the Mean Squared Error (MSE) loss function). The baseline technique can also surpass the performance of conventional ML models such as RF and GB. On the other hand, it can be observed that the baseline method is not robust enough when it comes to learning from the expanded feature set S.

A meta-learner estimates the weights of the ensemble learners in both techniques, but the combiner structure varies. It is important to note that the DCC-DNN method performs worse when using the SG method to combine the predictions of the correlations and the proposed DNN model. The drop in performance is due to the constant weights learned by these combinations. On the other hand, our implemented MoE method specializes in various feature spaces. It can adapt better to the new data by applying a conditional approach to decide the weighting of each predictor's output.

Figure 4.5 shows the predicted C_D generated by the proposed DCC-DNN as a function of Re and compared to other conventional methods and ML techniques. The data and predictions used to make the graphs were generated from the test split obtained from every fold of the appropriate cross-validation technique. We can appreciate that the DCC-DNN can predict more accurately C_D than other TC and ML-based methods for various ranges of ϕ. Especially in the case where $\phi \leq 0.5$, the proposed DCC-DNN can learn better the behavior of C_D than the other methods and predictions—the DCC-DNN also produces a smoother curve. Many TC approaches are restricted to certain Re and ϕ ranges based on the experimental data available at



(a) $\phi = 0.1$ (b) $\phi = 0.3$

(c) $\phi = 0.5$ (d) $\phi = 0.7$

(e) $\phi = 0.9$ (f) $\phi = 1.0$

Figure 4.5: Plot comparison of Re vs. C_D for different ϕ ranges. The data and predictions shown are from the test splits gathered from different folds of the applied cross-validation process. High-resolution color is recommended for the best viewing experience.

the time these methods were introduced. While our proposed DCC-DNN performs well in both high and low Re, TC techniques such as Yow *et al.* [162] and Holzer & Sommerfield [148] have difficulty predicting accurate C_D at higher Re. In the literature, tree-based machine learning approaches such as RF and GB have favored training a model on particle drag data, with several of these algorithms achieving encouraging results [152, 159]. However, our DCC-DNN demonstrated superior generalization capability, obtaining a better fit than RF and GB, particularly for $\phi = 0.1$ and $\phi = 0.3$. When dealing with noisy data, tree-based methods have also been susceptible to overfitting [166].

4.2 Multi-Source Weak Supervision Fusion

We propose a multi-source weak supervision fusion technique to train on a highly imbalanced dataset annotated with noisy labels. Using a Confident Learning technique, we lessen the effect of the noise while boosting the quality of the labels. We combine relevant predictions from models trained on large-scale visual datasets using Differential Evolution to better perform on underrepresented target features. In the TRECVID2021 Disaster Scene Description and Indexing (DSDI) Challenge, our technique achieved the top score among all the submitted runs, independent of the training data utilized.

Image and video recognition algorithms have advanced rapidly and with better precision and are expected to become a critical component of incident and disaster responses [167]. Using advanced technologies and deep learning methodologies such as Convolutional Neural Networks (CNNs), it is possible to deploy a drone ahead of the search team to identify the most damaged areas prioritized during a disaster swiftly. The automated content-based analysis and classification of the observed

disaster-related features in recorded videos will allow better curation and retrieval of critical information for situational awareness. Due to insufficient training data and standards, most existing methods do not fulfill public safety demands.

Civil Air Patrol (CAP) has the technical capability to function even when severe weather disrupts power, internet, phones, and airplane takeoffs, making it a critical and cost-effective tool for the Federal Emergency Management Agency (FEMA) to survey the impacted region swiftly and efficiently. CAP offers aerial pictures of flooded areas, collapsed dams, and other natural disaster-related events. To this end, several large-scale disaster imagery datasets, including the Incidents Dataset [168], LADI (Low Altitude Disaster Imagery) [18], xBD [17], etc., have been recently released to stimulate the development of new research and technologies in this field. Given the volume of data being collected, it is also critical to develop sophisticated tools and systems for curating all of the information.

It is challenging to analyze the images taken by low-altitude planes since they have a low height perspective, an oblique angle, and many disaster-related parts that image recognition systems do not usually take into account. We propose a weakly-supervised learning technique that incorporates data from various sources, many of which are of low quality or have been trained on subjects significantly different from the target classification task. The proposed fully-automatic solution would substantially decrease the time and expense associated with the classification jobs while delivering superior outcomes.

The main contributions of this paper are summarized as follows.

- We propose a new semi-supervised training technique robust to noisy, limited, and erroneous annotations and class labels from multiple sources.

- For the multi-source weak supervision fusion framework, a unique approach is proposed to recognize and merge the relevant predictions from various pre-trained networks.
- The proposed approach is evaluated on the LADI dataset and achieved the top score among all the submitted runs in the TRECVID2021 [21] Disaster Scene Description and Indexing (DSDI) Challenge, independent of the training data utilized.

4.2.1 Motivation

Images or video recordings may assist emergency responders in quickly inspecting the damages after a disaster event. Hence, there is a need for advanced techniques, such as Convolutional Neural Networks, to help curate and retrieve critical information at the desired moment. However, most existing approaches fail to fulfill public safety criteria due to a lack of suitable training data and standards. Most current solutions rely on high-quality annotations to build reliable models that can sufficiently automate image processing and concept detection. Non-experts are likely to have only seen low-altitude photos on rare occasions. Consequently, obtaining sufficient high-quality annotations to build a viable training dataset will be too expensive.

Most current solutions rely on high-quality annotations to build reliable models that can sufficiently automate image processing and concept detection. Non-experts are likely to have only seen low-altitude photos on rare occasions. Consequently, it will be too costly to get enough high-quality annotations to build a good training dataset. Numerous researchers have developed a variety of deep learning algorithms that are less reliant on the quality of the training data. The weakly-supervised tags and visual information train semantic-aware hash functions [169]. Previously, deep

canonical correlation analysis (DCCA) [170] was used to combine visual and text tag data. Many previously reported techniques rely on sparse line reconstruction, sparse coding, and dictionary learning to recover textual tags, which costs time and space and is not suited for large-scale applications. Research into automated disaster scene descriptions from images has grown in popularity. Newly-released disaster datasets such as xBD [17] and the Incidents Dataset [168] feature a top-down and a ground-level view of the damages. However, LADI [18] is unique in its images' low-altitude and oblique views. More recent studies explore an ensemble learning approach to tackle the class-imbalance and noisy-label issues [171, 172]. Incorporating spatio-temporal information to increase the model's contextual awareness has also been investigated [19]. Our proposed framework aims to improve the quality of noisy labels in the LADI training data through a Confident Learning (CL) [173] strategy. Furthermore, a novel multi-source information fusion method is proposed to improve the performance of the underrepresented target features in LADI.

4.2.2 Proposed Methods

Figure 4.6 illustrates the entire flow of our proposed framework. CL is used to improve the quality of the noisy labels in the crowdsourced annotated training set, which is the first step in our multi-source architecture for combining weak supervision from different sources. Several semantically related predictions are utilized to enhance the performance of a target feature using the scores of several semantic concepts derived from various machine annotators. The text that describes the target feature is turned into high-dimensional vectors, which are then used to look for semantic similarity and pick relevant concepts from other networks. We optimize

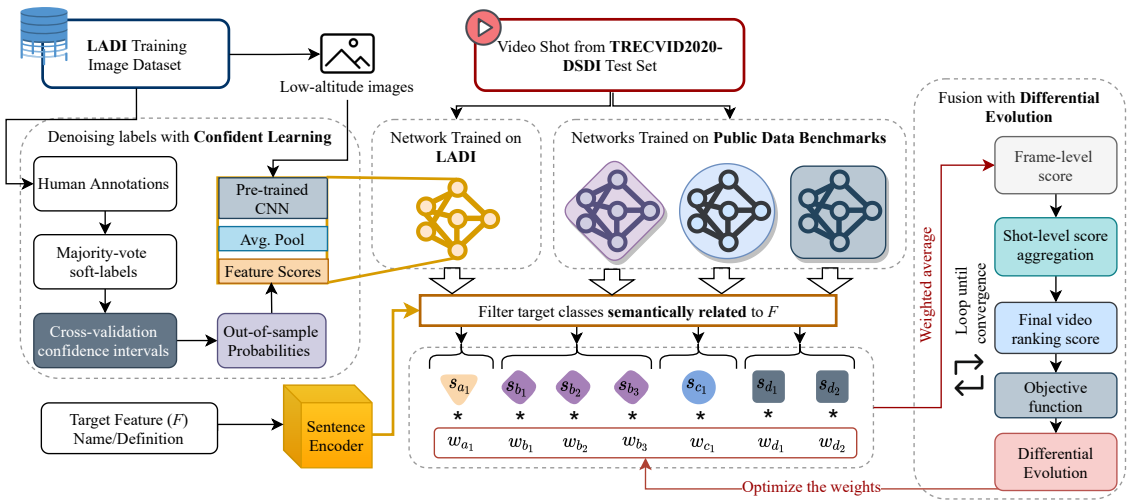


Figure 4.6: The proposed weakly-supervised deep learning framework implements a confident learning approach to denoise crowdsourced annotations and a multi-modality fusion framework to search and combine relevant target features predicted by multiple networks.

a weighted average that incorporates all of the models' relevant predictions into a single scalar that serves to rank the video clip using Differential Evolution (DE).

Denoising with Confident Learning

Naturally, we cannot do supervised classification without using labels related to our training data. Thanks to tools like Amazon Turk [174], crowdsourcing labels are now more accessible as a cost-effective method to parallelize complex annotation jobs and perform them quickly. This is due to the enormous number of Turkers, as well as the fact that MTurk is an on-demand service, with jobs being performed virtually immediately after they are posted. It's also worth noting that crowdsourcing removes various obstacles to admission and provides access to a wide collection of people with skill sets that would otherwise be unavailable.

According to the LADI researchers [18], annotations are organized as Human Intelligence Task (HIT), which asks the human worker whether any of the target

features in each of the five categories (i.e., *damage*, *environment*, *infrastructure*, *water*, and *vehicle*) are correct. Each HIT is allocated to up to five workers (asking just one category at a time) in order to reach an agreement on the label quality. Namely, for an image i and a target feature F_C that belongs to a specific category C (i.e., $F_C \in C$), the initial soft score S_{i,F_C} is calculated as follows.

$$S_{i,F_C} = \frac{\#Positive\ Votes_{i,F_C}}{Total\ Votes_{i,C}} \tag{4.10}$$

To calculate S_{i,F_C}, we assume that a particular image must have at least one vote from an annotator who was assigned a specific category C. Then, we employ cross-validation confident intervals [173] to derive out-of-sample prediction probabilities to improve the label quality further.

Multi-Source Weak Supervision Fusion

Machine Annotators: In this study, we employ four CNN network configurations (i.e., ResNet50, DenseNet161, YOLOv4, ViT-B/16, and LADI+Others) pre-trained on four open-source dataset (i.e., Places365, Incidents Dataset, MS COCO, ImageNet21k, InceptionV3). ResNet50 [38] and DenseNet161 are pre-trained on the Places365 dataset, which contains 1.8 million training images taken from 365 scene categories [175]. Another ResNet50 network is also pre-trained on the Incidents Dataset, containing 446,844 manually annotated images that cover 43 incidents across various scenes [168]. YOLOv4 (You Only Look Once) [176] pre-trained on Microsoft Common Objects in Context (MS COCO) is one of the leading deep learning-based object detection frameworks. The ViT-B/16 [177] model pre-trained on the ImageNet21K dataset is a key component in our proposed framework. Last but not least, an InceptionV3 model trained on LADI plus other sources by Presa-Reyes *et al.* [19] has also been incorporated.

Multi-Source Concept Fusion: Given the scores of different semantic concepts from various machine annotators, there might be multiple related ones that can be utilized to identify a target feature F. A Universal Sentence Encoder based on the Deep Averaging Network (DAN) [178] converts text describing the target feature into high-dimensional vectors T that are then utilized for semantic similarity from the cosine distance of the vectors. To fuse multi-source concepts, the high-dimensional vectors of the target feature F and the semantic concept P are first matched, and the weighted average score of those closely correlated concepts are fused, i.e.,

$$S_F(k, w_F) = \sum_{p \in O} w_F^p \cdot X_k^p \tag{4.11}$$

where $O = \{P|\theta(T_F, T_P) > \vartheta\}$, $w_F^p \in \mathbb{Q}$, and $0 < w_F^p < 10$. Furthermore, there exist multiple key frames inside a given video shot v. Therefore, the average score over all the key frames in v is computed as the shot-level feature score, which can be formally written as

$$S_V(V, w_F) = \frac{1}{||V||} \sum_{k \in V} S_F(k, w_F) \tag{4.12}$$

Then, for a given dataset of video shots $\mathcal{V}$ and a target feature F, the the top-N shot with F can be defined an ordered sequence $V_F = [V_1, V_2, \ldots, V_N]$, where $V_i \in \mathcal{V}$ and $\forall i > j, S_V(V_j, w_F) > S_V(V_i, w_F)$.

Weight Optimization based on Differential Evolution: The remaining problem is to determine the optimal weights w_F for each target feature F. Differential Evolution (DE) is a kind of evolutionary optimization technique that works with a population of candidate solutions. It uses genetic operators like mutation and recombination to repeatedly enhance the population. The objective function G determines each candidate's fitness. If $G(s_1) < G(s_2)$, candidate s_1 is judged superior to candidate s_2. The objective function seeks to improve the average precision

for a specific target feature (i.e., minimize $1 - AP^N$) by measuring the performance of a collection of retrieved results using the precision and recall metrics. Assuming the solution contains N video shots ordered by the final aggregated confidence score, our objective is to minimize the following error formula:

$$\hat{w}_F = \arg\min_{w_F} G(w_F) = \arg\min_{w_F} \left[1 - AP^N(V_F)\right] \tag{4.13}$$

The resulting weighted average combines all of the models' semantically relevant predictions into a single scalar that serves to rank the video clip based on a certain attribute.

4.2.3 Experimental Analysis

Table 4.2: Performance comparison among our proposed technique and competing methods.

Method	**Training Data**	**Precision@k**			**Recall@k**			**F1@k**			**MAP**
		k=10	**k=100**	**k=1000**	**k=10**	**k=100**	**k=1000**	**k=10**	**k=100**	**k=1000**	
BUPT_MCPRL [21]	L	0.271	0.225	0.228	0.271	0.232	0.405	0.271	0.227	0.244	0.159
VCL_CERTH [179]	L+	**0.510**	0.367	0.245	**0.511**	0.415	0.378	**0.510**	0.377	0.255	0.282
Presa-Reyes *et al.* [19]	O	0.413	0.392	0.285	0.413	0.448	0.682	0.413	0.404	0.316	0.298
Ours-CL-BA	L	0.394	0.346	0.279	0.394	0.383	0.648	0.394	0.351	0.307	0.254
Ours-CL-ZS	O	0.471	0.409	0.296	0.471	**0.522**	0.789	0.471	0.425	0.332	0.339
Ours-CL-DE (proposed)	L	0.384	0.351	0.286	0.384	0.395	0.683	0.384	0.360	0.315	0.268
	O	0.481	**0.425**	**0.310**	0.481	0.502	**0.793**	0.481	**0.439**	**0.345**	**0.359**

L+ LADI-based (L) training data plus additional human annotations (i.e., instance and segmentation).

Dataset

We test our methods using the LADI dataset, which comprises images acquired by CAP from a low-flying aircraft and maintained by FEMA. The LADI training dataset consists of images captured from an airplane, and the LADI test dataset consists of brief video clips captured from a UAV. The DSDI track's test dataset

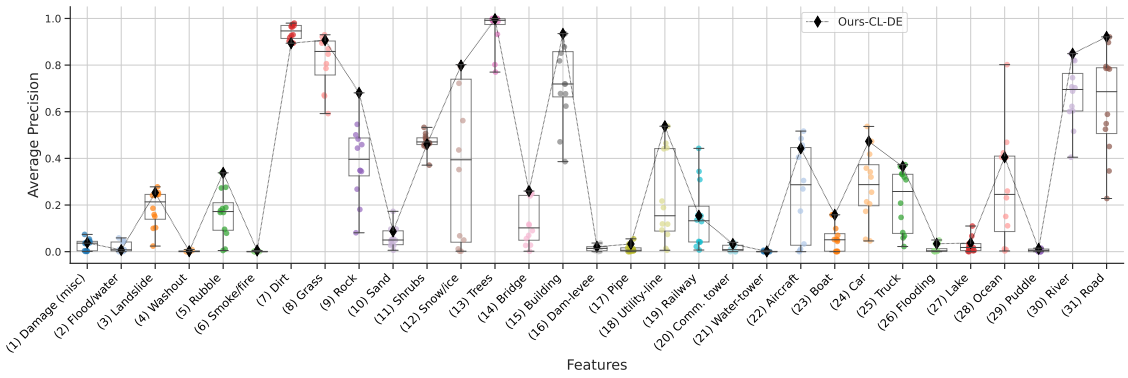


Figure 4.7: The boxplot shows the distribution for a feature's precision score compared across all submissions to TRECVID2021-DSDI, independent of which training dataset was used to train each technique. The placement of our proposed method's performance is demonstrated using a black diamond.

in 2021 comprises 2,802 video shots with a maximum duration of 60 seconds per shot, focusing on the devastation wrought by an earthquake tragedy. The test set supplied in TRECVID2020-DSDI is used as validation during the DE processing in our case. The Mean Average Precision (MAP) metric is used to examine and compare the performance of different approaches.

Competing Methods

To determine the effectiveness of the proposed technique, we compare it to other competing methods, such as BUPT_MCPR [21] and VCL_CERTH [179]. Both competing methods are trained solely on the LADI dataset. In particular, the problem was approached by VCL_CERTH as a panoptic segmentation problem. They provide their own instance and semantic segmentation annotations for 300 LADI images to train the panoptic network. The method proposed by Presa-Reyes *et al.* trained on LADI plus other datasets was also included. Two baseline fusion techniques using the average of the best performing model (Ours-CL-BA) and aggregated Z-scores (Ours-CL-ZS) are also explored to compare to our proposed DE fusion.

Feature Score Model and Fusion

Two feature score models, EfficientNet-B5 [180] and ResNet50 [38], are trained on the LADI's confident labels generated by the CL-based approach. Using transfer learning, we fine-tune the network's weights on ImageNet. The network's final classification head is replaced with a thick layer using sigmoid activation for multi-class soft-label classification. With a starting learning rate of ($\eta = 1e-4$), we use the Adam solver to optimize our model. For the differential evolution search, we employed the DE/best/1/bin technique, which generates new candidate solutions by randomly picking solutions from the population, subtracting one from the other, and adding a scaled version of the difference to the population's best candidate solution.

Results and Discussion

The proposed framework is compared to competing methods mainly categorized as LADI-based (L) the LADI + Others (O) track submission—where "Others" in our proposed approach refers to the inclusion of models pre-trained on open-source data benchmarks. Table 4.2 summarizes the performance comparison across different methods. The excellent results obtained by the panoptic segmentation approach proposed by VCL_CERTH on the LADI-based (L) track underline the necessity to integrate additional information about the images other than the noisy labels.

Our proposed technique achieves impressive results on the LADI + Others (O) track, particularly compared to other competing methods. The high recall rate illustrates our classification model's ability to detect and recover the majority of positive examples within a relevant target feature. By comparing the baseline methods Ours-CL-BA and Ours-CL-ZS, we demonstrate our proposed Ours-CL-DE approach can find better weights when aggregating the predictions of different models. Further-

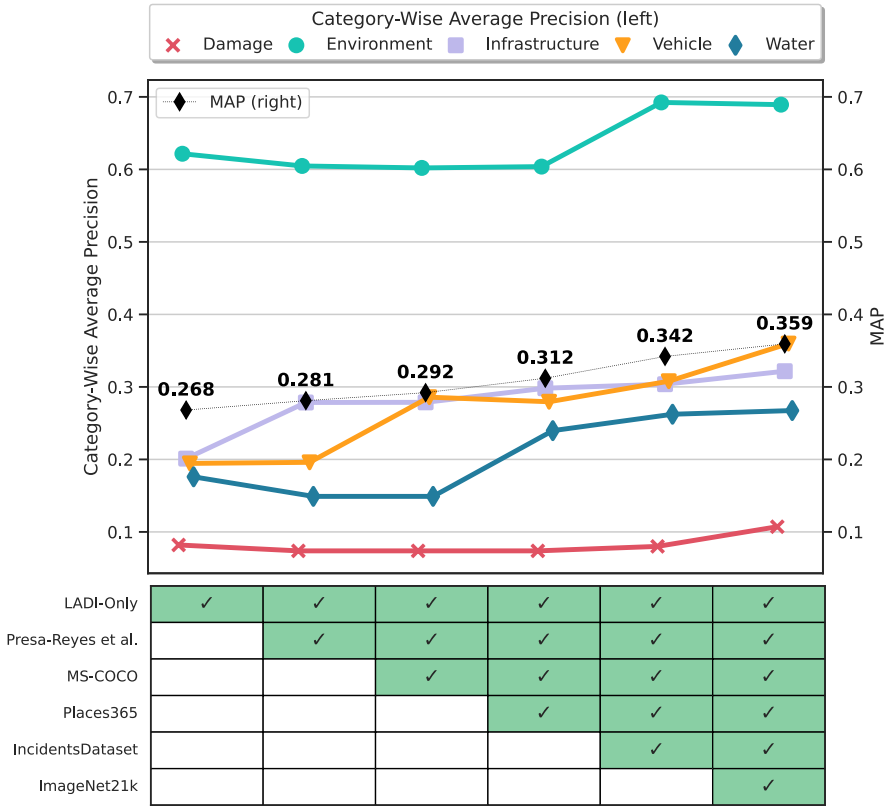


Figure 4.8: Optimal weights learned by the proposed multi-source weak supervision fusion technique of semantically-relevant concepts.

more, compared to the method that simply trains on the LADI-based (L) data, the proposed method introduced on LADI + Others (O) exhibits a considerable improvement of about 34% in the MAP score, indicating the effectiveness of our strategy of fusing the weak supervision from multiple sources.

In Figure 4.7, the average precision at the target feature level shows that our suggested approach has obtained the greatest performance for the target features such as debris, rock, snow/ice, building, utility-line, boat, river, and road. Figure 4.8 depicts the performance contribution of each additional dataset used to train the machine annotators.

Starting from our proposed CL technique trained on LADI only, each additional dataset is added to the ensemble as depicted by a checkmark in the figure. The ResNet50 pre-trained on the Incidents Dataset contributed a performance boost for the environment category, detecting concepts such as 'snow covered' and 'field' and improving on features of snow/ice and grass. The damage features, on the other hand, did not improve as expected given the damage concepts from the Incidents Dataset, necessitating further investigation. The environment features achieve better performance because they are simpler to discern from long distances and show lower inter-class variation than other categories. YOLOv4 network pre-trained on MS COCO contributes a performance boost for the vehicle categories, detecting concepts for 'aeroplane,' 'boat,' 'car,' and 'truck.'

We employ a weighted average ensemble that achieves better performance thanks to the integration of human and machine-generated annotations. Since it is clear that the relationship between the relevant features is not linear to their semantic similarity, our proposed technique has been proven to be a viable approach to identifying the best predictions based on the performance of each machine annotator. Because our proposed technique outperforms existing methods with minimal training, they are an excellent means of leveraging and transferring information from the methods that have already been presented in previous research into any emerging topic.

Table 4.3 qualitatively summarizes the top 10 video clips retrieved for six of our best performing target features. This qualitative visual is meant to help compare the performance among our proposed methods trained on LADI only; our previously submitted method from last year trained on LADI + Others; and our proposed improvements trained on LADI + Others. Our proposed method is demonstrated to work best when the prediction from multiple models is available. Our technique

Table 4.3: Qualitative results comparing the first 10 video clips retrieved for six features using three of our submitted solutions, namely, Ours-CL-DE-L (top-row), Presa-Reyes *et al.* (middle-row), and Ours-CL-DE-O (last-row).

works well in retrieving suitable clips for target features like car or road, even if their visual attributes are widely diverse. Furthermore, by applying the transformer techniques such as ViT, the proposed approach is able to recognize smaller items in a picture, such as the utility-line target feature.

4.3 Conclusion

Due to the technological advances in artificial intelligence and machine learning, now more than ever, multi-modal and multi-source data fusion has the potential to deliver enhanced tools for several applications in fluid dynamics and disaster situational awareness. This chapter presents a unique approach to developing a general drag model that combines scientific knowledge of previously specified empirical correlations with the prediction capacity of the neural networks. We introduced an expanded feature set that captures key elements of the single-particle free-fall experiment, such as flow property, particle geometry, and settling direction. Along with the usual Reynolds number Re and sphericity ϕ found in the literature, we further include density ratio R_ρ, aspect ratio AR, crosswise sphericity $\phi_\perp$, and lengthwise sphericity $\phi_\parallel$ to enhance the modeling method. We developed a novel DNN model architecture trained to predict C_D using our defined expanded set of features as input. The proposed method uses several regularization techniques to address data issues such as limited experimental data, extreme data points, and noisy information. We further use the meta-learning method to combine the predictive capabilities of the DNN model with the scientific knowledge of the previously proposed empirical correlations to develop the DCC-DNN framework. A ten-fold cross-validation method was used to construct folds that adhere to the constraint of non-overlapping experimental sources. Compared to earlier models, DCC-DNN predicted the C_D of

spherical and non-spherical particles with tremendous accuracy and excellent capability to adapt to various experimental scenarios. According to the RMSE measure, our proposed methods were capable of decreasing the error to about 30% in comparison to the performance of the baseline. We envision DCC-DNN as a reliable and inexpensive technique for industrial-scale design and application.

Furthermore, most present-day picture recognition algorithms fail to meet public safety requirements due to a lack of appropriate training data. As part of our multi-source weak supervision fusion architecture, we apply the CL technique to enhance the quality of noisy labels in the crowdsourced annotated training set. Semantic similarity is used to identify relevant concepts predicted by other networks. We use DE to rank the video clip based on a weighted average of all relevant model predictions. Combining many classifiers pre-trained on well-known data standards improves the overall performance. Still, only the best and most relevant predicted score for a particular target feature should be used. Overall, the study shows how this framework has great potential to save a significant amount of time and resources while still achieving outstanding results in the disaster scene description task. Although this work focuses on disaster scene description, the proposed methods have been developed with extendability. Our approaches effectively leverage and transfer knowledge from past studies into any new topic. As a potential future work, we will explore more advanced techniques of incorporating other multi-modality sources using our proposed method, such as spatio-temporal data.

CHAPTER 5

FINE-LEVEL PATTERN MODELING WITH DEEP FEATURE FUSION

When there are significant areas in the data from two or more incredibly similar classes, this is known as class overlap or inter-class similarity. A high inter-class similarity makes distinguishing between the two or more classes difficult, if not impossible. Other problems, such as class imbalance, are made more difficult by overlapping characteristics among minority and majority classes. Mining the fine-level patterns from the data is necessary to identify the fine-grained features that can effectively discriminate across different classifications.

Several researchers have shown that extracting features from only a few critical portions of the input data may be utilized to infer more discriminating information than extracting features from the entire data set [51, 181]. A relevant characteristic (global or local) may also include valuable discriminative information that enables one item to be differentiated from another. For instance, local features describe image patches (small groups of pixels), whereas global features explain how the whole image is composed of pixels. In this chapter, we can derive local features that automatically guide our suggested fine-level pattern modeling approach toward identifying honey bee subspecies by taking into account domain knowledge data. Furthermore, we show how, at many layers of the deep learning architecture, we can detect and fuse fine-level patterns from the data, addressing both local and global interactions of the data's features to create a more accurate prediction for complex samples to tell apart.

5.1 Processing Local Fine-Level Patterns

Fine-level pattern modeling aims to identify acceptable characteristics among visually similar groupings [182]. With successful applications to the geometric morphometrics-based approach in identifying honey bee subspecies, we further demonstrate the extendibility of the presented method. Geometric morphometrics is a set of techniques for describing biological forms mathematically using geometric descriptions of their size and shape [109]. These morphological features are nodes at the junction of wing veins that relate to landmarks for honey bees. These wing morphological characteristics have effectively distinguished honey bee species and subspecies. In some instances, the approach may be as precise as more complex mitochondrial techniques [110]. However, annotating these morphometric characteristics requires the observer to be familiar with each species, an understanding that a novice is not expected to have. The preliminary results demonstrate a remarkable performance of 90.82 average precision using a Keypoint R-CNN ResNet-50 Feature Pyramid Network (FPN) [183] to detect the morphological features automatically. We extract local features that can automatically guide our proposed technique to identify the honey bee subspecies by interpolating the landmarks into a structured grid.

5.1.1 Motivation

Honey bees are essential for food security since their pollination services enable the high production of many fruits and other crops such as apples, almonds, melons, broccoli, blueberries, and others. According to a recent assessment, the estimated economic value of pollination services globally is $351 billion. Because certain HB species and subspecies have unwanted features, it is critical to distinguish between them. Apis mellifera scutellata, for example, was mistakenly introduced into Brazil

in 1956. It has spread over much of South and Central America and some southern US states. It crossed with the European A. m. ligustica employed by beekeepers as it expanded, resulting in a feral Africanized population that is more aggressive than the European strain it surpassed.

Humans and animals have died due to this species' enhanced protective behavior in the United States and across Central and South America. Another HB species with negative traits is A. m. capensis, native to Africa's southern cape. Its ability to reproduce parthenogenetically allows it to parasitize colonies of its geographic neighbor, A. m. scutellata, eventually killing them and negatively impacting its beekeeping industry. Tools that can distinguish between insect species are desperately needed. Because of the influence of insects on human health and agriculture, this is critical. For example, it is crucial to distinguish between insect species that transmit harmful infections like viruses to people and agricultural plants and those that do not. Unwanted insect species may cause problems to crops, causing financial losses and jeopardizing food security.

5.1.2 Landmark Identification

Morphological characteristics have long been employed in entomology to identify insects, and honey bees are no exception. The digitization of insect collections has had a beneficial effect on geometric morphometric analysis methods, a tool for identifying insects utilizing face-recognition algorithms based on critical geo-referenced locations in a given canvas to assign IDs to particular groups. We created a high-resolution method to identify honey bee species and subspecies by combining morphological and molecular data using AI algorithms. This data will be combined with genetic

and behavioral profiles and environmental and ecological data to confirm subspecies and species assignments.

The model Keypoint R-CNN ResNet-50 Feature Pyramid Network (FPN) [184] was chosen from pre-trained models. Keypoint R-CNN is an extension of the well-known Mask R-CNN model, designed for semantic and instance segmentation. In Mask R-CNN, a Region Proposal Network (RPN) first applies a basic CNN architecture to extract feature maps to obtain bounding box candidates from the input image. These candidates represent the presence of objects. The model is then trained to identify the bounding box and segmentation masks closest to the object of interest. Mask R-CNN models the keypoint's location as a one-hot mask for each K keypoint type for keypoint detection.

5.1.3 Subspecies Classification

Fine-level pattern modeling aims to identify the subtle characteristics to discriminate amongst superficially similar groupings. When there are significant areas in the data from two or more highly similar classes, class overlap, also known as inter-class similarity, arises. Differentiating between two or more classes is difficult, if not impossible, due to considerable inter-class resemblance. Other problems, such as class disparity, are worsened by the overlapping characteristics of the minority and majority classes. It is critical to extract fine-grained patterns from data to discover additional features that may help the model differentiate between different classes. Previously, fine-level pattern modeling methods could only extract hand-crafted features from input data while ignoring layout information.

To a non-expert, the images of the honey bee's wings from different species are almost indistinguishable. Deep neural networks can capture many complicated

characteristics that humans cannot understand. At multiple layers, the extracted feature maps are convolved by 1×1 convolution that follows works as a coordinate-dependent transformation implemented to input the extracted feature vectors and define the correspondence at each spatial location (x, y). This convolution approach leads to dimensionality reduction, with its combination being mathematically equivalent to a multi-layer perceptron at each (x, y) [185].

5.1.4 Experimental Analysis

Experimental Setup

The images used were about 1600x1200 pixels in size each. Images were divided into groups of training (80%) and validation (20%). The model is trained on the training set, and the validation set corroborates how well the model performs on unseen data. Due to the limited size of the dataset and the close semblance between the images, data augmentation techniques (including random rotations, flips, changes in brightness, contrast, and saturation) were applied to the training dataset to avoid overfitting the model [186].

Landmark Identification with Keypoint Detection

We present an application of a state-of-the-art AI model trained to automatically identify 19 diagnostic morphological features of HB wings from images. The model's performance demonstrates the capability of successfully applying the latest AI keypoint detection techniques to the study of geometric morphometrics for the discrimination of HB species or subspecies. Our current tests show that the transfer learning strategy was successfully applied for both high and poor quality images, enabling us to develop accurate models in a time-saving manner.

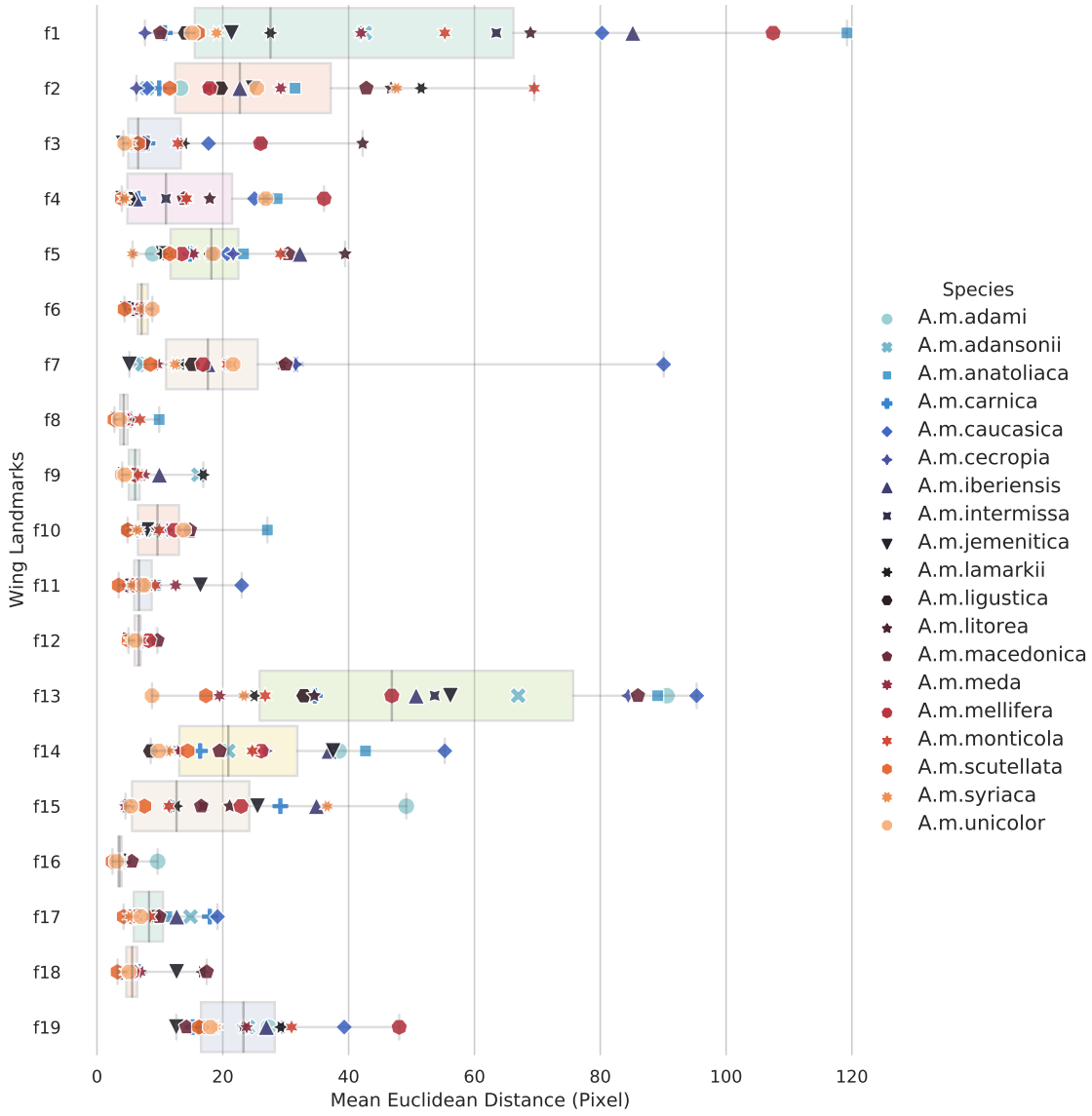


Figure 5.1: Been wing landmark identification performance comparison grouped by species for each landmark.

Figure 5.1 illustrated the performance comparison grouped by species for each landmark. For the 19 species used for validation, 16 of the 19 points were successfully identified with less than 2% error. Moreover, 18 of the 19 points were correctly recognized with less than 3% error. The points with the highest and lowest average error are 1 and 19. The species with the highest average error rate was A. m. ligus-

tica with 2.9%, and the lowest was A. m. ruttneri with 0.8%, and the average error rate across all 15 species was 1.8% (Figures 6 and 7). To improve the keypoint identification further, we will review more advanced techniques to identify the 19 points, including deeper, more complex models such as Keypoint R-CNN ResNet-101 FPN. In addition, we will further automate the morphometric process by developing a second step to discriminate the wings among different species or subspecies, provided the keypoints are identified and projected into a common coordinate space.

After the model was trained for 102,500 iterations, inferences were made against the testing set to compare the results' ground-truth labels. The preliminary results demonstrate remarkable performance for keypoint detection (Table 4), calculated by the Object Keypoint Similarity (OKS) metric [187], showing how close the predicted keypoint is to the true keypoint. OKS is determined by dividing the distance between expected and ground truth points by the wing's scale. The formula for OKS is as follows:

$$OKS = \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \tag{5.1}$$

Where: d_i is the Euclidean distance between the ground truth keypoint and predicted keypoint; s is the scale calculated as the square root of the wing's segmented area (i.e., the overall area that contains the keypoints); and k is a constant per-keypoint that controls the fall-off.

All points are initialized with a minimal k value of 0.025 to train the model to identify the keypoint as closely as feasible to the real keypoint. Given OKS, the Average Precision (AP) is calculated using three well-known thresholds in the literature. Mainly it serves to compare the OKS value with the given threshold. If OKS is greater than the threshold, then the keypoint is detected.

Table 5.1: Performance summary of the proposed keypoint detection method on the validation split.

Metric	Performance
AP[OKS=0.50:0.95]	90.821
AP[OKS=0.50]	98.884
AP[OKS=0.75]	96.872

Table 5.2: Performance summary of the proposed subspecies classification method on the validation split.

Method	Accuracy	Precision	Recall	F1
Baseline	0.67	0.68	0.64	0.65
Proposed	0.74	0.73	0.72	0.72

Bee Subspecies Identification using Local Fine-Level Patterns

Discriminating information enables one item to be differentiated from others by processing the significant characteristics (global or local). However, global features have disadvantages, including noise sensitivity, illumination changes, scaling, and the inability to identify the image's essential elements [60]. Table 5.2 demonstrates the comparison between a baseline method and the proposed local fine-level pattern modeling. The confusion matrix in Figure 5.2 shows each bee subspecies' performance using the proposed method.

5.2 Deep Feature Fusion for Two-Stream Networks

The selection and combining of features in feature fusion techniques are concerned with eliminating duplicate or superfluous traits in the data. To generate superior hybrid features, a good feature fusion technique should take advantage of correlations while retaining a complete set of distinct traits. We propose a two-stream CNN architecture that solves the challenges of identifying buildings at four damage

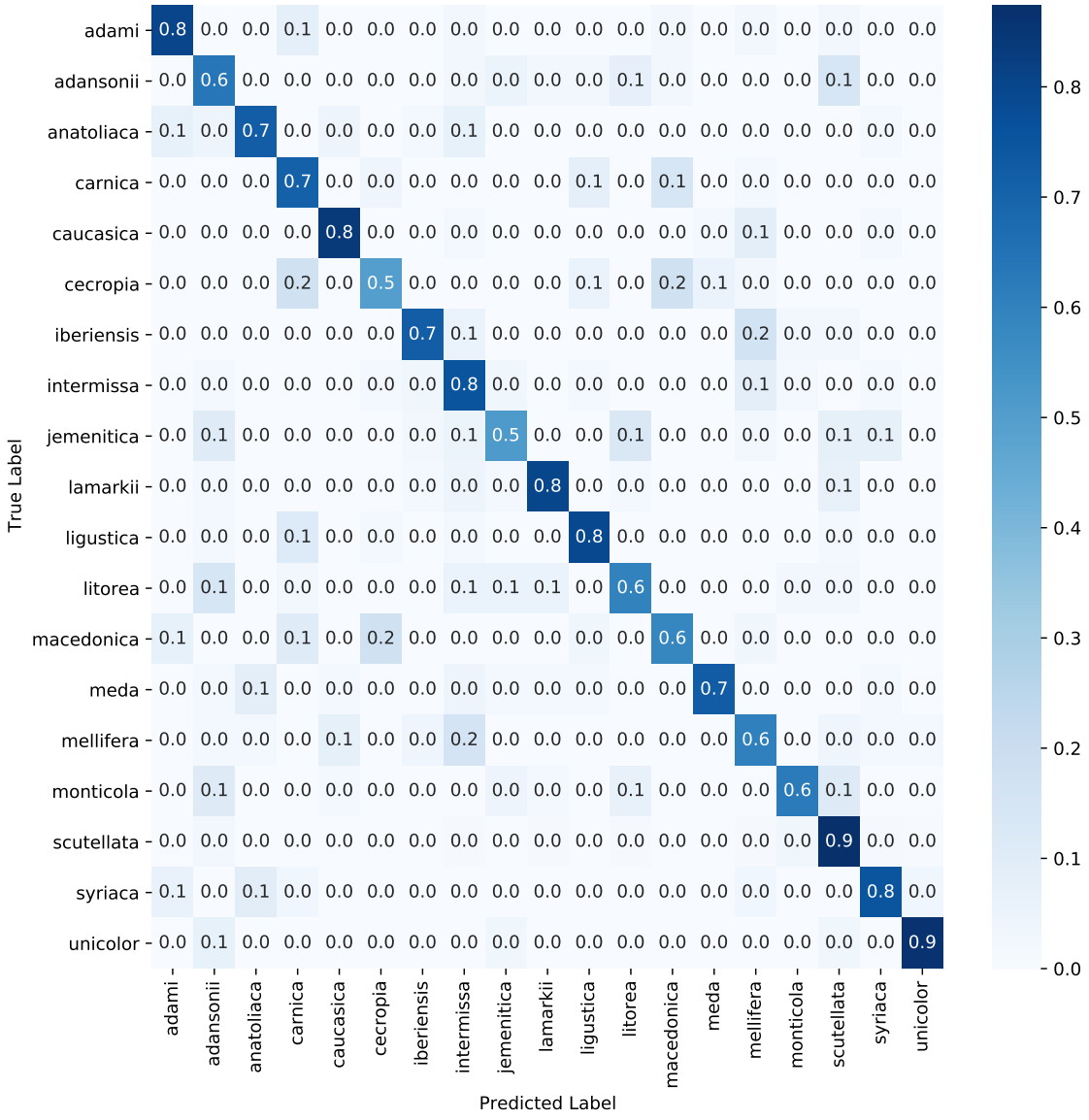


Figure 5.2: Confusion matrix of the honey bee subspecies identification results on the validation split.

levels. We test it using a curated, fully labeled dataset compiled from accessible sources. We aim to combine the characteristics produced by the two streams so that the total model can understand how the inputs from both networks relate.

Remote sensing data, which is the data collected by a high-flying aircraft or the satellite scanning of the earth, has become a crucial tool to rapidly survey the

damage of the affected regions with limited access. Current efforts have started to rely on aerial photographs captured after the tragedy to assess the resulting damage and make decisions [188]. However, searching and assessing the damage concerning specific regions becomes a laborious and time-consuming job for emergency responders when large areas are affected. Hence, there is a need for tools and methods to aid the rapid assessment of damaged homes in these time-sensitive situations.

Machine learning and deep learning have shown tremendous success in various research areas [189, 190, 191, 1]. Convolutional Neural Networks (CNNs) are renowned for achieving state-of-the-art performance in image classification. Given the current technological developments to acquire massive volumes of data and the recent advances in artificial intelligence and machine learning, now more than ever, disaster information integration and fusion have the potential to deliver enhanced situational awareness tools for humanitarian assistance and disaster relief efforts [192]. We propose a model trained on pre- and post-disaster satellite imagery. As a case study, we selected the event of Hurricanes Irma and the damage it has inflicted on the Florida Keys.

5.2.1 Motivation

The technological advances in data collection and the growing availability of high-quality data have enabled substantially automated and rapid damage assessment research. Indeed some of our proposed techniques were inspired and improved upon Fujita *et al.*'s research previously published in 2017 [10]. The study proposed applying a two-stream CNN model to classify whether a building survived or was washed away by a tsunami that followed the wake of the Great East Japan earthquake on March 11, 2011. Another source of inspiration and information was the survey con-

ducted after Hurricane Irma by Xian *et al.* [193], assessing the damage to more than 1600 residential buildings in the Florida Keys. The study also provided a statistical analysis that reveals distinct factors potentially influencing the degree of damage in some of the most affected regions.

Ci *et al.*, for example, presented a novel technique by integrating CNNs for feature extraction and a new loss function called ordinal regression for optimizing classification results to analyze the extent of building damage caused by the 2010 Yushu and 2014 Ludian earthquakes [194]. Moreover, in the Summer of 2019, the Defense Innovation Unit (DIU) announced the xView2 Challenge [17] which aimed to stimulate applied research focusing on automating assessing building damage after a natural disaster. DIU publicly released a high-quality and large-scale dataset known as xBD, composed of satellite imagery annotated with building localization and levels of damage before and after natural disasters.

The previously introduced damage assessment studies work well when classifying damaged buildings into two categories (i.e., intact or destroyed). Our proposed approach aims to overcome the difficulties of categorizing the building at different damage levels and significantly boost the model's performance.

5.2.2 Data Preprocessing

Our study finds that the surroundings of a building provide critical contextual information and visual cues that can help the model better predict the building's level of damage. However, including too much of the surrounding area increases the risk of confusing the model by inadvertently feeding it image patches containing multiple buildings of different damage levels. As shown in Figure 5.3, after resizing the bounding box to a size 80% larger than the footprint's geometric bounding box

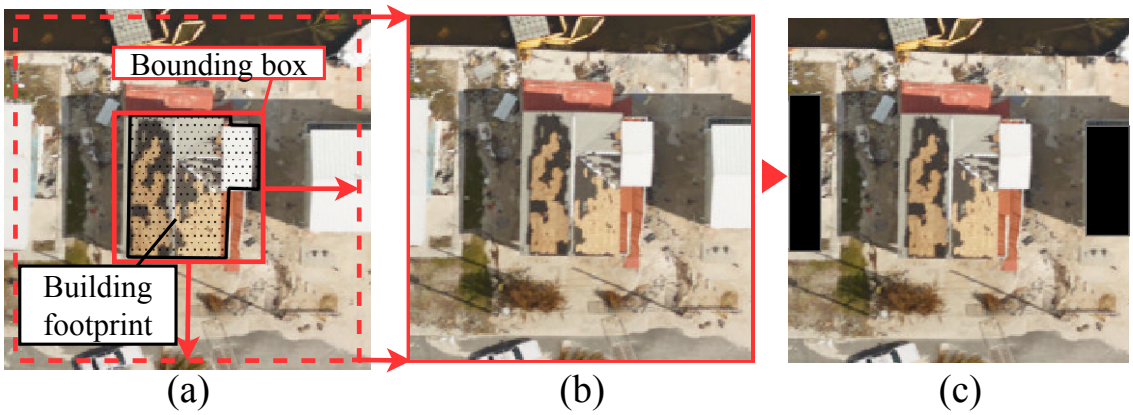


Figure 5.3: The input patch preprocessing steps start with (a) the bounding box surrounding the building's footprint being extended to cover enough surrounding area; (b) then the resized patch containing the building in the center is cropped, and (c) finally, nearby buildings are occluded to avoid confusing the model.

and cropping the widened patch containing the building in the center, the nearby buildings are occluded by negating the pixels found inside the surrounding buildings' footprints. Hence, for each input pair, the model will only focus on the building in the center, and the clues found in the surrounding area will improve the performance. This method is very effective in reducing the uncertainty of the model.

5.2.3 Feature Extraction

Figure 5.4 demonstrates the proposed architecture with both CNN streams following the ResNet50 architecture pre-trained on ImageNet [38]. The last classification layer is removed to fuse the outputs of the networks' average pooling layer and fine-tune the weights of the entire network. Both networks are trained on their unshared weights, giving each the flexibility to learn specific features from their individual input data stream. Training both streams on unshared weights has shown to be much more effective than sharing the weights [10], especially when there is a significant gap between the images' time, causing the image's appearance to be distinct.

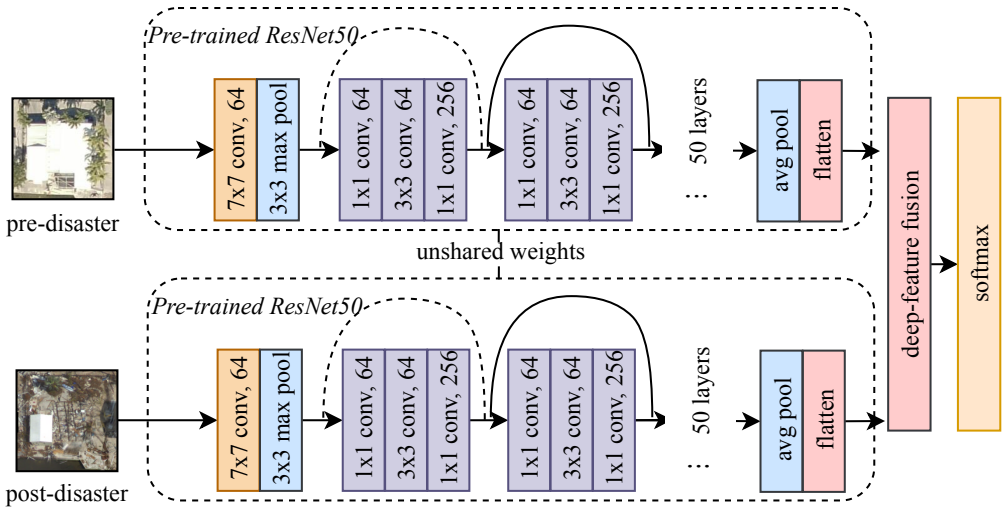


Figure 5.4: The proposed two-stream CNN architecture.

5.2.4 Deep Feature Correspondence

Our goal is to fuse the features generated by the two streams to allow the model to learn the correspondence of the inputs from both networks [43]. Namely, a fusion function $f : \mathbf{x}_n^l, \mathbf{x}_n^r, \rightarrow \mathbf{y}_n$ fuses the feature pair at the output of both models' n-th layer given the feature maps $\mathbf{x}_n^l \in \mathbb{R}^{H \times W \times D}$ and $\mathbf{x}_n^r \in \mathbb{R}^{H' \times W' \times D'}$ to produce the output feature $\mathbf{y}_n \in \mathbb{R}^{H'' \times W'' \times D''}$, where H, W, and D constitute the height, width, and number of channels of the corresponding feature map. In our approach, f is applied as a late-fusion. The following assumptions are made, $H = H' = H''$, $W = W' = W''$, $D = D' = D''$, and the subscript n is dropped for simplicity purpose.

Concat Fusion. $\mathbf{y}^{\text{cat}} = f^{\text{cat}}(\mathbf{x}^l, \mathbf{x}^r)$ joins the feature sequence along an existing axis. In our case, both features are joined across their width:

$$\mathbf{y}^{\text{cat}} = \text{concat}\{x_{i,j,d}^l, x_{i,j,d}^r\},$$

where $1 \leq i \leq H$, $1 \leq j \leq W$, $1 \leq d \leq D$ and $\mathbf{y} \in \mathbb{R}^{H \times 2W \times D}$. Concatenation is one of the most common fusion techniques. This type of fusion often works well and is simple to implement. However, this technique does not define a correspondence between the features.

Conv Fusion. $\mathbf{y}^{\text{conv}} = f^{\text{conv}}(\mathbf{x}^l, \mathbf{x}^r)$ first the two feature maps are stacked at the same spatial location (i, j) across channel d, namely:

$$y^{\text{stack}}_{i,j,2d} = x^l_{i,j,d}, \qquad y^{\text{stack}}_{i,j,2d-1} = x^r_{i,j,d},$$

where $\mathbf{y} \in \mathbb{R}^{H \times W \times 2D}$. The corresponding layers define the correspondence by learning the suitable filters when convolving the stacked data.

$$\mathbf{y}^{\text{conv}} = \mathbf{y}^{\text{stack}} * \mathtt{f} + \mathtt{b},$$

where $\mathtt{b} \in \mathbb{R}^{D'''}$ is the term for bias and $\mathtt{f} \in \mathbb{R}^{1 \times 1 \times 2D \times D'''}$ is a bank of filters. In our approach, the total number of filters is set to $D''' = 100$. Thus, $\mathtt{f}$ is trained on the weighted combinations of $\mathbf{x}^l$ and $\mathbf{x}^r$ at the same feature location.

5.2.5 Experimental Analysis

Dataset

For Hurricane Irma, Monroe County released a preliminary damage assessment report [105]. Although the assessment appears to be incomplete, only including references for some of the majorly damaged and destroyed residential homes, it still provides a good reference on how to get started with labeling the different types of damages. The official Monroe County survey evaluation was integrated with Xian *et al.* [193] data to create the dataset summarized in Table 5.3. The following lists the

definition of each classification label according to FEMA's official guide to assessing damage and impact [108].

Table 5.3: The statistical information of the building damage dataset.

No.	Concepts	# of Instances
1	No damage	17,482
2	Minor damage	2,633
3	Major damage	962
4	Destroyed	1,436
	Total	22,513

Experimental Setup

The fusion techniques (i.e., `concat fusion` and `conv fusion`) are compared with a scenario (called `post only`) where one pre-trained ResNet50 model is trained on the images taken after the disaster event. Our approach is evaluated on our fully-labeled dataset depicting various damage levels resulting from Hurricane Irma's landfall on the Florida Keys. Our inputs are also divided into two sets (1) `crop center`, the size of the patch is only as large as the size of the bounding box surrounding the building footprint, the information in the patch will focus mainly on the building rooftop, and the final patch sizes are 112×112; and (2) `extended patch`, where the bounding box is extended to cover enough surrounding area as described in Sec. 5.2.2, and the final patch sizes are 224×224.

Data augmentation is applied to enhance the training set to overcome the highly-imbalanced nature of the Hurricane Irma dataset (as depicted in Table 5.3). Data augmentation is a common approach that improves the performance of the CNN models and its generalization capability by applying random transformations to the input data. The augmentation techniques randomly applied to our dataset consist of horizontal and vertical flips, rotation, shear transformations, and zooming.

Moreover, given the limited size of the dataset, 5-fold cross-validation is utilized to evaluate our model and prevent over-fitting. 80% of the images are randomly chosen for training and 20% for testing in each fold. On each fold, each model configuration is trained for 80 epochs.

The Adam solver is employed to optimize our model with an initial learning rate ($\eta = 0.0001$) that is small enough to update the transferred weights slowly when fine-tuning the model and achieve a more optimal set of final weights [195]. During training, the learning rate will drop to 10% of its current learning rate after no improvements to the testing loss for 10 epochs. Moreover, the model is also trained using small batches of size 16. The smaller batch sizes introduce noise to the training process, leading to a regularizing effect that improves the model's generalization capability for a given computational cost [196].

Experimental Results

Table 5.4: Results summary of each class's F1-score averaged from the results of 5-fold cross-validation.

Config	**Crop Center**			
	no-dmg	**minor-dmg**	**major-dmg**	**destroyed**
Post-Only	0.938	0.673	0.530	0.747
Concat Fusion	0.956	0.745	0.587	0.819
Conv. Fusion	0.957	0.749	0.616	0.834
Extended Patch				
Post-Only	0.951	0.739	0.582	0.815
Concat Fusion	0.969	0.819	0.701	0.875
Conv. Fusion	**0.971**	**0.831**	**0.711**	**0.901**

Table 6.5 demonstrates the performance summary for the class-specific F1-scores under the three configurations previously described as well as considering the `crop center` and `extended patch` input scales. The results report the average of five runs

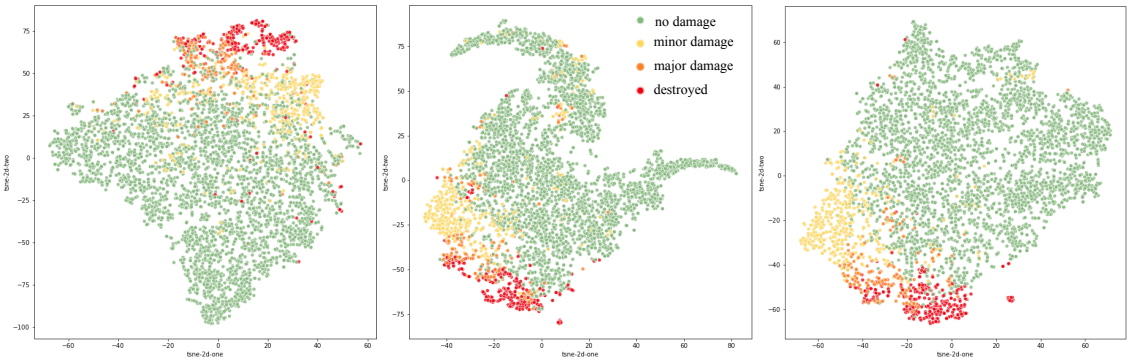


Figure 5.5: t-SNE visualization of the feature layer right before the softmax layer for the first fold in cross-validation trained on the input data of the extended patch for (a) post only; (b) concat fusion; and (c) conv fusion

from the cross-validation setup, with a standard deviation ranging from $\sigma = 9.6\text{e}{-04}$ to $\sigma = 3.8\text{e}{-02}$. The experimental results demonstrate the effectiveness of our proposed approach, which (1) `extended patch`, pre-processing the data by extending the building patch to include enough surrounding context while occluding surrounding buildings of potentially different damage levels; and (2) `conv fusion`, applying an advanced fusion technique by learning the correspondence between two feature maps. The proposed approach consistently achieves the best results throughout the four damage categories.

Various aspects emerged when analyzing the results. First, the fusion of deep features from the pre-post image pairs performs better than the `post only` configuration, even though there is a significant gap of 2 years between the times both images were captured. Besides the time difference, there are also substantial changes in angle, illumination, and resolution. Moreover, because the imagery must be captured promptly after the event, it is common for post-disaster photos to have clouds present and sometimes present blurriness. The performance improvements demonstrate the need to include a reference of how the building looked before possibly

being affected by the event to make a better assessment. Second, there is also a boost in performance by extending the image patch to include the building's surroundings (i.e., `extended patch`) compared to using the image patches by including mainly the rooftop information (i.e., `crop center`).

The configuration `concat only` is a prevalent approach to fusing the features and performs well when compared with the `post only` option. However, it is demonstrated that the model's performance can be improved by considering the correspondence between the features. Figure 5.5 shows the capability of each model configuration to separate the four categories using t-Distributed Stochastic Neighbor Embedding (t-SNE) [197] to visualize the features from the layer right before the softmax classifier. t-SNE is a powerful tool that can map high-dimensional data into a lower dimension and has an insight into how the data is arranged. While Figures 5.5(a) and 5.5(b) demonstrate some overlaps between different damage levels, Figure 5.5(c) shows a better separation of the classes.

Table 5.5: Performance measures of the proposed network configuration.

Damage Type	**Precision**	**Recall**	**F1-Score**
No Damage	0.96	0.98	0.97
Minor Damage	0.87	0.79	0.83
Major Damage	0.73	0.69	0.71
Destroyed	0.90	0.90	0.90

Our proposed model obtains a weighted F1-score of 0.94, calculated by first computing the F1-score for each damage class, then finding their average weighted by the number of true instances in each class, considering the imbalanced nature of the data. Table 5.5 summarizes the Precision, Recall, and F1 scores on a class-by-class basis for the proposed approach. The model demonstrates the weakest performance when classifying major damaged instances due to the limited number

of samples available (as shown in Table 5.3) and the overlaps between major and minor damaged buildings.

5.3 Multi-Level Convolutional Fusion and Graph-Transformer Networks

We propose a novel deep learning architecture to automatically assess the building damages from remote sensing data comprising satellite imagery across various disaster types and damage levels. By integrating the features captured from pre- and post-images at multiple levels, the proposed approach solves well-known challenges in categorizing buildings at four damage levels. Our proposed method is evaluated on a large-scale building damage assessment dataset. Our approach considers the building as a single entity, demonstrating greater accuracy while making predictions that would be easier to interpret.

5.3.1 Motivation

Disasters introduce severe disruptions to a community's usual activities, exceeding its capacity to manage the damages using its resources. It is critical to make the appropriate judgments and take action in the affected area, given a thorough understanding of the situation, including the damage's locations, causes, and severity. More than ever before, emergency responders can quickly and remotely undertake a comprehensive damage assessment utilizing remote sensing survey techniques such as satellite images and aerial pictures. However, manually identifying damaged structures may be time-consuming and difficult when disasters hit a vast region. As

a result, gathering raw, high-quality images is insufficient to offer in-the-moment situational awareness.

A preliminary damage assessment is an on-site investigation of damages or failures caused by accidents or natural disasters. These damage assessments record the extent of damage that can be replaced, restored, or reclaimed. It may also estimate how long repairs, replacements, and rehabilitation will take. Recent advancements in disaster information processing and fusion can promote situational awareness during critical circumstances. Currently proposed methods in automatic building damage assessment using satellite images can be categorized into two approaches, instance-level damage classification [10, 194, 198, 9] and semantic segmentation [199, 200, 201]. Recently published studies in a building damage assessment study focused on semantic segmentation algorithms that yielded promising results. However, the pixel-level predictions may be unclear in the real-world application of such an approach.

Satellite images deliver a high-quality overview of the Earth's surface, making it a crucial tool for emergency responders to rapidly assess the damages in regions that may become inaccessible due to blockages of the roads or damages to communication infrastructures. In recent years, deep learning has enabled image processing research to grow fast, with remarkable results in image recognition, object detection, and semantic segmentation [1]. It is no surprise that deep learning techniques applied to satellite imagery are becoming increasingly prevalent as a research subject with many promising real-world applications.

Gueguen & Hamid introduced a semi-supervised damage detection framework to combine the advantages of previous supervised and unsupervised methods by requiring less labeled data and simultaneously achieving greater detection accuracy [202]. By proposing the simultaneous use of pre- and post-tsunami photos in a single Con-

volutional Neural Network (CNN), Fujita *et al.* [10] was one of the first studies to introduce the notion of a two-stream architecture in the context of building damage assessment. Other efforts have also opted to model the damage assessment as an ordinal classification [194]; however, it can be proven difficult to establish the sequence-relation within the classes when different types of damage are present in the dataset. xView2 has recently released xBD, a large-scale dataset created for the advancement of building identification and damage assessment across multiple damage levels and disaster types [17]. In the xView2 data competition, several teams from across the globe competed and improved on the baseline method [17]. Valemtijn *et al.* [198] and Nex *et al.* [203] concentrate on the practicality of CNN-based methods and how they perform on xBD data when used in operational emergencies, taking into account unseen data and time restrictions.

Semantic segmentation-based building damage assessment attempts to identify each pixel for a particular input image, whether or not it belongs to a building [204, 205, 206, 201] and what level of damage the building sustained. Semantic segmentation assigns a class label to each pixel in an image. Multiple objects of the same class are treated as a single entity. On the other hand, instance segmentation considers several objects of the same class as separate instances. However, in practice, evaluating individual buildings is a more convenient process.

Most of the recently published works in building damage assessment research have processed the xBD dataset at the pixel-level, achieving promising results when evaluating such methods using the pixel-level metrics. However, the pixel-level predictions may be incomprehensible in the real-world implementation of such a strategy. For instance, these pixel-level predictions make it challenging to answer critical queries like how many buildings sustained a specific amount of damage. Answering such a question may assist emergency management in making more accurate

evaluations of the current situation, enabling them to deploy resources where they are most required efficiently. Hence, the proposed approach aims at assessing the damages of each building at an instance-level while also aiming to produce accurate predictions.

5.3.2 Proposed Methods

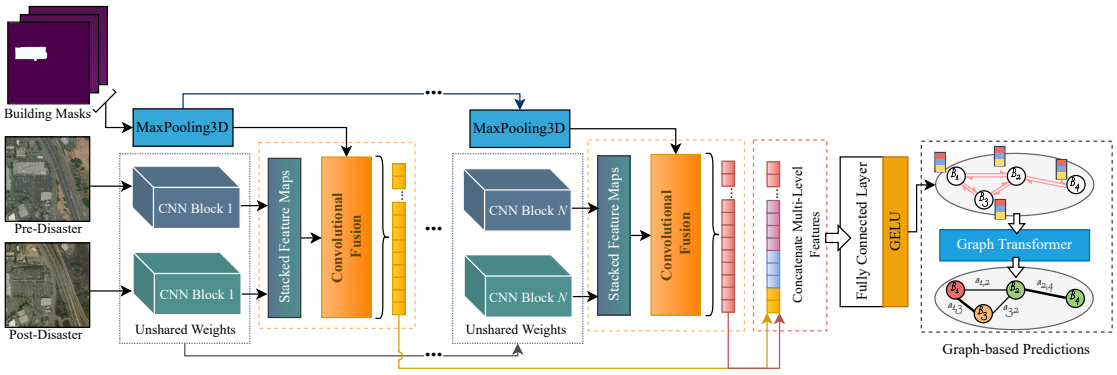


Figure 5.6: The proposed instance-level building damage assessment architecture.

As summarized in Figure 5.6, the proposed instance-level building damage assessment architecture takes as input a pair of satellite image tiles captured before and after a disaster occurrence together with an assembly of binary masks encoding the polygon representing the target building's area within the tiles. The proposed multi-modality fusion is implemented throughout multiple levels of the architecture, capturing both local and global feature interactions. We further enhance the model's performance by applying GNNs to propagate the retrieved multi-level features across the graph to nearby buildings.

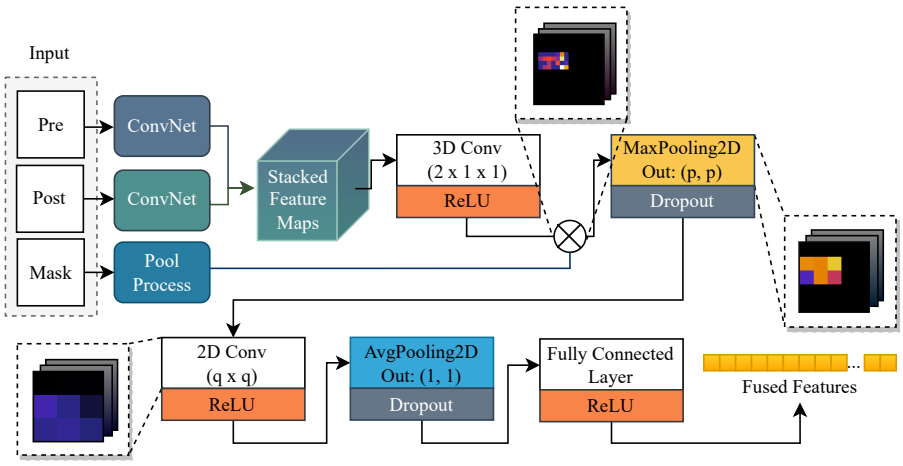


Figure 5.7: The proposed fusion component for the instance-level building damage assessment framework.

Multi-Level Convolutional Fusion for Feature Correspondence

The goal is to integrate the deep features provided by the two CNN streams so that the model can learn how the inputs from both networks correspond [9, 15]. As a result, CNN streams do not share weights, giving each feature-extraction branch of the model the ability to explore the most significant features between pre- and post-disaster satellite images and fuse them to capture their interactions. A fusion function is proposed, where $f(\mathbf{x}^{\mathtt{pre}}, \mathbf{x}^{\mathtt{post}}) \rightarrow \mathbf{y}$ combines the feature maps at an specific layer from both CNN models—i.e., $\mathbf{x}^{\mathtt{pre}}, \mathbf{x}^{\mathtt{post}} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels in the individual feature map. The fusion functions f are applied at the last layer of the convolutional processing blocks. Our proposed fusion function uses a three-dimensional convolutional layer to learn the feature correspondence between two streams.

$$\mathbf{y}^{\text{conv}} = \mathbf{y}^{\text{stack}} \times \mathtt{f} + \mathtt{b} \tag{5.2}$$

where $\mathtt{b} \in \mathbb{R}^{C'}$ is the term for bias and $\mathtt{f} \in \mathbb{R}^{C' \times 2 \times 1 \times 1}$ is a bank of filters. The subsequent layers specify the correspondence by learning the appropriate filters while convolving the stacked data. Hence, $\mathtt{f}$ is trained on the weighted aggregation of the feature maps $\mathbf{x}^{\mathbf{pre}}$ and $\mathbf{x}^{\mathbf{post}}$. The fused visual feature maps are multiplied with the set of building binary masks that have been processed in parallel by a max-pooling layer to a consistent size as the feature map to filter the features that correspond to each target building, as shown in Figure 5.7. The adaptive max-pooling layer further emphasizes the features that belong to a particular building while outputting a (p, p) feature map—the value for p ranges from 5 to 1 depending on the layer L from which the feature maps are extracted. Earlier layers will output wider feature maps, necessitating higher p. A convolutional layer further refines the building-specific features applying filters of (q, q) in which $q = \lceil p/2 \rceil + 1$. Finally, to construct the final fused features, an average pooling layer computes the average of the components present in the feature map area, followed by a fully-connected layer.

The surroundings of a building provide vital contextual information that may assist the model in better estimating the building's damage level [9, 10]. Therefore, the proposed convolutional fusion is applied at multiple levels of the architecture, capturing both local and global feature interactions. As the CNN streams process the pair of image tiles, the features related to the relevant building remain intact. Earlier layers aid in identifying finer-scale characteristics, such as small broken parts of roofing after a hurricane. Later layers better represent a building's surroundings, increasing the model's capacity to identify the damages related to disasters such as floods.

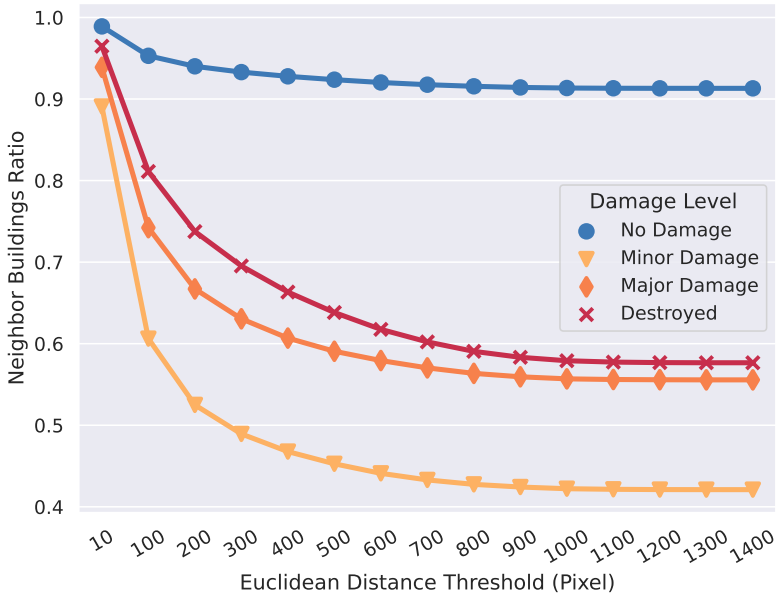


Figure 5.8: A plot illustrating the change of the likelihood that the neighboring building(s) within a certain distance of a specific building has the same damage level as the distance threshold increases.

Nearby Building Feature Fusion with Graph Transformers

One technique to express spatial data through graphs in spatial analytics is using spatial weights. They've often used constructs for representing geographic links between observational units in a spatially linked dataset. Geographical weights implicitly relate items in a geographic table to one another based on their spatial connections. Spatial weights are the critical method by which the spatial links in geographical data are brought to bear in the following analysis by articulating the idea of geographical closeness or connectedness. Waldo Tobler proposed the first rule of geography in 1970 [207], stating that "everything is typically connected to everything else, but those close to each other are more related than those that are further apart."

Figure 5.8 illustrates a trend that is in line with Tobler's rule. Except for structures that sustained minor damage, most of the surrounding buildings inside the threshold are likely to have suffered the same damage level even as the euclidean distance threshold rises. When it comes to buildings that suffered minor damage, the likelihood will drop to less than 50% when the threshold reaches less than or equal to 260 pixels—this distance ranges from 325ft to 819ft considering that the Ground Sampling Distance (GSD) in xBD training set is reported to range from 1.25ft to 3.15ft.

In image classification, CNNs take patch-wise input and produce a collection of one-shot encoded predictions. Unlike CNNs, GNNs need an adjacency matrix to input samples into the network. The majority of GNN algorithms need the creation of an adjacency matrix before commencing the training process. We denote a graph as $\mathcal{G} = (V, E)$, where V represents the nodes in the graph capturing the individual building's features and E represents the edges capturing the spatial relations among different buildings. We make use of the Graph Transformer [208] to propagate the multi-level extracted features across $\mathcal{G}$ to the connected buildings. This approach employs a gate mechanism trained to combine only the most relevant features. Given node features $H^l = \left\{h_1^l, h_2^l, \ldots, h_n^l\right\}$ of the l-th layer, we firstly calculate the attention weighted adjacency matrix between entity node h_i^l and its neighbors $\mathcal{N}^i$ in $\mathcal{G}$ as follows.

$$\mathbf{A}_{c,ij}^l = \frac{\left\langle \mathbf{Q}_{c,i}^l, \mathbf{K}_{c,j}^l + e_{c,ij} \right\rangle}{\sum_{u \in \mathcal{N}^i} \left\langle \mathbf{Q}_{c,i}^l, \mathbf{K}_{c,u}^l + e_{c,iu} \right\rangle} \tag{5.3}$$

where $\langle q, k \rangle = \exp\left(q^T k / \sqrt{d}\right)$ represents the exponential scale dot-product function, and d is the head's hidden size. Moreover, $\mathbf{Q}_{c,i}^l = W_{c,q}^l h_i^l + b_{c,q}^l$ is the query vector, and $\mathbf{K}_{c,j}^l = W_{c,k}^l h_i^l + b_{c,k}^l$ is the key vector. The edge features e_{ij} are encoded as

$e_{c,ij} = W_{c,e}e_{ij} + b_{c,e}$ and added into key vectors as additional information for each layer.

After obtaining the multi-head graph attention, we apply an aggregation of messages from the neighbor building node j to the relevant building node i.

$$\hat{h}_i^{l+1} = \|_{c=1}^{C} \left[\sum_{j \in \mathcal{N}^i} \mathbf{A}_{c,ij}^{l} \left(v_{c,j}^{l} + e_{c,ij} \right) \right] \tag{5.4}$$

where $\|$ is the concatenation operation for C head attention and $v_{c,ij} = W_{c,v}^{l} h_j^{l} + b_{c,v}^{l}$ encodes the target feature h_j for a weighted sum. We also employed the gated residual link between layers to prevent the model from over-smoothing and enhance the generalization capability of the model [208].

Constructing the Loss Function to Handle Class Imbalance

As shown in Table 3.1, the xBD data is highly skewed, in almost more than 70% of the buildings remained unaffected after a disaster event. Equation 5.5 computing the proposed heuristic weighting mechanism was inspired by the dynamic sampling mechanism proposed by Pouyanfar *et al.* [209]. The objective is to generate a set of heuristic weights $\hat{\mathbf{w}} = (\hat{w}_0, ..., \hat{w}_{K-1})$ for K classes where $\hat{\mathbf{w}} \in \mathbb{R}^K$.

$$\hat{\mathbf{w}} = \frac{N}{n_c} \times \frac{1 - \mathtt{F1}_c}{\sum_{c_k \in C} (1 - \mathtt{F1}_{c_k})} \tag{5.5}$$

where N represents the total number of training samples, n_c represents the number of samples for a particular class c, and $\mathtt{F1}_c$ represents the reported state-of-the-art performance for each class c.

Other than class imbalance, another major challenge is the overlapping features found within some of the damage classifications. The Categorical Cross-Entropy (CCE) loss function [210] is used to evaluate the model's performance.

The Focal Loss (FL) [211] function is utilized in conjunction with the suggested class-weighting technique to aid the model in making a more accurate distinction between difficult-to-distinguish classes. Focal Loss down-weights well-classified instances and concentrates on the challenging examples. For a sample misclassified by the classifier, the loss value is much higher than the loss value corresponding to a well-classified case. For each integer label $y \in \{0, ..., K-1\}$ from class K and $\hat{\mathbf{p}} = (\hat{p_0}, ..., \hat{p}_{K-1}) \in [0,1]^K$ is a vector of an estimated probability distribution over the K classes.

$$\mathrm{CCE}(y, \hat{\mathbf{p}}) = -\log(\hat{p}_y) \tag{5.6}$$

$$\mathrm{FL}(y, \hat{\mathbf{p}}) = -1 \times \alpha_y \times (1 - \hat{p}_y)^\gamma \log(\hat{p}_y) \tag{5.7}$$

The FL function modifies the classic cross-entropy loss by introducing a modulating factor $(1-\hat{p}_y)^\gamma$, with an adjustable focusing parameter $\gamma \geqslant 0$—our proposed mechanism sets $\gamma = 2$ and $\alpha_y = \hat{w}_y$. The combined loss function $\mathcal{L}$ between FL and CCE is calculated as follows.

$$\mathcal{L} = ((1-\beta) \times \mathrm{CCE} + \beta \times \mathrm{FL}) \times \log(R) \tag{5.8}$$

where $\beta = 0.2$ and R is the total pixel-area of the relevant building footprint. Class-weighting works well with the proposed loss function to emphasize the most difficult samples to recognize. Moreover, $\log(R)$ puts more weight on the bigger buildings, helping the model prioritize the buildings that occupy a larger area and achieve better performance on the pixel-level metrics.

5.3.3 Experimental Analysis

Dataset and Evaluation Metric

xBD [17] is used as a benchmark to assess the performance of our methods; it is a recently introduced large-scale dataset created for the advancement of building identification and damage assessment across multiple damage levels and damage types. xBD offers pre- and post-event multi-band satellite imagery with building polygons, classification labels for damage forms, ordinal damage level labels, and corresponding satellite metadata from several disaster events. There are bounding boxes for environmental variables such as fire, water, and smoke in the dataset. The dataset contains around 700,000 building annotations across over $5,000$ $\texttt{km}^2$ of imagery from 15 countries.

The details of the xBD data are summarized in Table 3.1. DIU made tier 1, and tier 3 data splits accessible at the start of the contests. Competitors did not have access to the ground-truth annotations of the xBD's test split; however, results could be uploaded to an online leader-board system to measure a model's performance and improve accordingly. The test split was used as a validation set. At the end of the competition, the best-performing models are further compared by utilizing a holdout set. Hence, our proposed project utilized tier1 and tier3, which make up 80% of the overall data, to train the model; the xBD test split was utilized as the validation set and the hold-out set as an unbiased test of the final trained model.

Due to the competition's chosen pixel-level metrics, semantic segmentation has been the preferred approach for submitting competing methods and previously proposed building damage assessment research aiming to achieve the best results on the xBD dataset. On the other hand, a semantic segmentation technique may have the drawback of not being able to answer interpretable and realistic queries essen-

tial for effective damage assessment operations. This study investigates semantic segmentation as a potential building localization solution, but with the added goal of distinguishing the distinct buildings. Hence, we make use of the xView2 Challenge metric [17], a weighted average of the F1 score of the building segmentation, and the harmonic mean of class-wise damage classification F1 scores defined by equation 5.10. The F1 measure is defined as follows, as shown in equation 5.9.

$$\text{Damage F1} = \frac{n}{1/(\text{F1}_{C_1} + ... + \text{F1}_{C_n})} \tag{5.9}$$

$$\text{xView2 Score} = 0.3 \times \text{Localization F1} + 0.7 \times \text{Damage F1} \tag{5.10}$$

The Localization F1 score is a binary F1 measurement of the performance of pixel-based localization predictions of buildings. Using the xView2 metric, our result is compared to the xView2's first place (XFP) team solution. Despite the significant differences in training approach between our solution and XFP, our study aims to make a fair comparison by training our localization method and comparing our proposed architecture to the performance of the single-trained damage assessment model, using the pre-trained weights shared by the XFP team [212]. Comparing both methodologies quantitatively and qualitatively demonstrates the reliability of our proposed instance-based damage assessment methodology in delivering more accurate and interpretable predictions.

Experimental Setup

Localization Model: The building localization based on satellite images can be regarded as an image segmentation task, where all the pixels that belong to the object building are recognized. The semantic segmentation model performs a binary classification task and determines whether the pixels in the image belong to a

building or not. Specifically, the U-Net architecture [213] is adopted for robust and accurate image segmentation, which has been explored in some previous works [214]. The U-Net architecture comprises two main components: contraction (i.e., encoder) and decoder. The former converts the input image into a latent representation, and the latter maps the latent representation to the pixel-level segmentation results. Among all the variants of U-Net architecture, in this paper, we adopt the U-Net with ResNet50 backbone to localize the buildings in the images. To achieve better model performance and accelerate the model training, the ResNet50 pre-trained on ImageNet dataset [215] is used to initialize the parameters of the encoder in the model. The rest of the model parameters are initialized by a warm-up training strategy [216], which freezes all the parameters of the encoder and optimizes the model based on the training dataset. The complete model is trained against the training dataset to achieve the optimal building localization performance. In both cases, the Adam optimizer [217] is applied to minimize the segmentation results [211]. A noise filtering process is applied on the predicted binary masks where the detected buildings with tiny sizes are ignored. The size threshold is determined based on the building size statistics of the dataset. We naively assume that the connected pixels that belong to a building are of the same building. Thus, the polygons representing each building in the image are computed using the boundaries of each connected region of building pixels.

Input Preprocessing and Data Augmentation: Albumentations [218] allows us to compose image transform operators in a variety of ways to enhance model performance. Composition enables the sequential application of numerous augmentations to an input picture or the use of basic control-flow logic. Each transformation in a composition uses the result of the preceding transformation as an input. This simple yet effective method allows you to design complicated transform

Table 5.6: Augmentation methods were applied during model training to both the training and validation splits of the dataset using the albumentations python library.

Augmentation Method	**Applied Split**	**Non-Default Parameters**	**Prob.**
Rotate 90°	Train Valid	-	20%
Sized BBox Safe Crop	Train Valid	erosion_rate=0.3 (T) erosion_rate=0.1 (V)	100%
Horizontal Flip	Train Valid	-	50%
Vertical Flip	Train Valid	-	50%
One Of	Train	-	10%
· Motion Blur		-	20%
· Median Blur		blur_limit=3	10%
· Blur		blur_limit=3	10%
One Of	Train	-	20%
· CLAHE		clip_limit=2	50%
· Sharpen		-	50%
· Emboss		-	50%
· Brightness and Contrast		-	50%
Hue Saturation Value	Train	-	10%

Table 5.7: Performance comparison of some instance-based building damage classification trained on the xBD dataset. Instance-level scores are reported, followed by the pixel-weighted score inside the parentheses.

No.	Method	Class-wise Damage F1				Damage F1
		no-damage	minor-damage	major-damage	destroyed	
1	xBD Baseline [17]	87.2 (86.4)	43.0 (32.3)	37.6 (37.5)	60.9 (59.8)	51.5 (46.6)
2	Valentiji *et al.* [198]	93.1 (93.8)	60.9 (50.5)	65.5 (70.9)	77.0 (82.6)	72.2 (70.6)
3	Presa-Reyes & Chen [9]	93.7 (94.3)	61.5 (54.9)	66.4 (71.0)	79.4 (84.6)	73.3 (73.1)
4	Proposed	**95.2 (95.6)**	**69.6 (65.9)**	**72.9 (77.4)**	**83.0 (87.6)**	**79.0 (80.1)**

pipelines that may be implemented in a variety of low-level array manipulation tools. A composition enables a simple and straightforward declarative description of such a complicated series of augmentations. The augmentation methods applied are summarized in table 5.6.

Implementation Details: Training is done for at most 40 epochs, and the best network is selected and stored based on its performance, measured via a macro-average of the validation F1 score, treating each class equally. In these experiments, the pre-trained model weights are fine-tuned using the Adam optimizer [217] with an initial learning rate ($\eta = 1e-4$). In addition, the learning rate is multiplied by a factor of 0.1 during training after no changes were observed to the validation loss for over ten consecutive epochs.

Results and Discussion

Results of the damage assessment model compared to the most recent instance-based building damage classification and an official baseline are described in Table 5.7. The performance of related research is evaluated by training the competing method's proposed networks on the data-split investigated in this study to provide a fair comparison with the proposed network. F1 metrics are reported at the instance level and by a pixel-weighted score included inside the parentheses.

Table 5.8: Quantitative comparison among different backbones and the introduction of the graph transformer. The best-aggregated F1-Scores are shown in bold for validation and test set.

Backbone	Graph Trans.	Class-wise Damage F1								Macro F1		Harm. F1	
		No Damage		Minor Damage		Major Damage		Destroyed					
		Test	Valid	Test	Valid	Test	Valid	Test	Valid	Test	Valid	Test	Valid
ResNet34		94.8 (95.1)	96.1 (95.7)	63.6 (57.5)	67.1 (60.2)	71.4 (77.0)	76.1 (77.2)	81.7 (86.6)	83.0 (87.0)	77.9 (79.0)	80.5 (80.0)	76.2 (76.3)	79.2 (77.7)
	✓	95.3 (95.4)	96.2 (95.5)	65.2 (61.1)	68.0 (60.3)	70.5 (76.6)	76.1 (78.4)	82.4 (86.6)	84.6 (87.9)	78.4 (79.9)	81.2 (80.5)	76.7 (77.8)	79.9 (78.2)
DualPathNet92		95.2 (95.5)	95.9 (95.7)	67.0 (60.9)	66.1 (59.4)	72.6 (77.0)	75.6 (77.4)	82.8 (87.3)	82.6 (86.7)	79.4 (80.2)	80.1 (79.8)	78.0 (77.9)	78.6 (77.3)
	✓	95.4 (95.6)	96.4 (95.9)	68.5 (62.2)	68.4 (61.4)	72.3 (76.7)	76.4 (78.4)	83.9 (86.9)	84.6 (88.2)	80.0 (80.3)	81.4 (81.0)	78.7 (78.3)	80.1 (78.7)
SENet-154		94.6 (95.2)	95.8 (95.6)	57.2 (53.7)	62.5 (60.2)	70.4 (75.5)	74.6 (76.0)	82.5 (87.4)	84.0 (87.9)	76.2 (77.9)	79.2 (79.9)	73.6 (74.3)	77.3 (77.5)
	✓	95.4 (95.9)	96.2 (95.6)	65.8 (62.1)	69.2 (62.6)	70.4 (76.5)	76.5 (75.6)	83.9 (88.1)	85.5 (88.6)	78.9 (80.7)	81.8 (80.6)	77.2 (78.5)	80.6 (78.5)
DenseNet201		95.3 (95.6)	96.3 (95.9)	62.3 (56.6)	67.1 (58.1)	71.6 (77.4)	75.7 (77.2)	82.9 (87.2)	83.9 (87.7)	78.0 (79.2)	80.8 (79.7)	76.1 (76.2)	79.3 (77.0)
	✓	95.5 (95.7)	96.4 (96.0)	68.6 (62.2)	69.4 (62.0)	72.4 (76.6)	75.8 (77.8)	83.7 (87.8)	84.1 (87.8)	80.0 (80.6)	81.4 (80.9)	78.7 (78.4)	80.2 (78.7)
SKResNeXt-50		95.2 (95.4)	96.1 (95.6)	64.3 (59.6)	67.5 (59.9)	71.9 (76.4)	74.8 (76.0)	82.4 (86.8)	83.3 (87.3)	78.4 (79.6)	80.4 (79.7)	76.8 (77.1)	79.0 (77.3)
(32x4d)	✓	95.5 (95.7)	96.2 (95.5)	67.1 (61.6)	69.5 (62.6)	71.3 (76.5)	76.3 (77.9)	83.8 (87.9)	84.4 (87.9)	79.4 (80.4)	81.6 (81.0)	77.9 (78.2)	80.4 (79.0)
EfficientNet-B5		95.2 (95.5)	96.1 (95.5)	66.9 (59.7)	65.4 (57.9)	72.1 (77.0)	73.3 (76.0)	82.0 (87.3)	83.2 (86.2)	79.0 (79.9)	79.5 (78.9)	77.6 (77.4)	77.9 (76.2)
	✓	95.5 (95.7)	96.3 (96.0)	69.5 (63.0)	66.7 (60.9)	73.0 (77.1)	75.1 (76.9)	83.5 (87.5)	85.0 (87.6)	**80.4** (80.9)	80.8 (80.3)	**79.1** (78.9)	79.3 (78.0)
SEResNeXt-50		95.1 (95.0)	96.1 (95.5)	69.3 (62.9)	67.5 (60.8)	74.2 (78.5)	75.3 (77.1)	81.9 (85.9)	83.9 (87.9)	80.1 (80.6)	80.7 (80.3)	79.0 (78.7)	79.3 (78.0)
(32x4d)	✓	95.2 (95.6)	96.3 (95.8)	69.6 (65.9)	69.3 (63.2)	72.9 (77.4)	77.0 (77.8)	83.0 (87.6)	84.7 (88.0)	80.2 (**81.7**)	**81.8** (**81.2**)	79.0 (**80.1**)	**80.6** (**79.2**)

Ablation Studies: We undertake comprehensive experiments using the following three aspects to better identify the advantages of various components of our suggested model:

- **Different architectures**: Similar to the XFP solution, our model tests the performance of the proposed approach with four pre-trained networks, namely, ResNet34, DualPathNet92, SENet-154, and SEResNeXt-50 (32x4d). Furthermore, three more pre-trained backbones are also explored; SKResNeXt-50 (32x4d), DenseNet201, and EfficientNet-B5, as summarized in Table 5.8. The architecture with the best performance, SEResNeXt-50 (32x4d), was chosen based on the validation score. SEResNeXt is a ResNeXt model variant that uses squeeze-and-excite blocks [219] to allow the network to undertake dynamic channel-wise feature recalibration.
- **Graph transformer**: The proposed aggregation of adjacent features using graph transformers demonstrates consistent improvements in Table 5.8 at both the instance-level and the pixel-weighted harmonic F1 score ranging from 0.9% (0.64%) up to 4.28% (2.4%) for validation set, and from 0.85% (0.43%) up to 4.97% (5.65%) for the test set.
- **Distance threshold and graph construction**: We measure the impact of the distance threshold utilized to construct the adjacent building spatial graph on the proposed framework's performance. Figure 5.9 summarizes the change in performance under different thresholds. It's worth noting that the distance criterion that reaches the best performance for both the validation and test sets (i.e., 260 pixels) is consistent with the maximum distance at which minor-damage buildings are no longer the majority of nearby buildings, as previously illustrated in Figure 5.8.

Figure 5.9: Performance impact is based on the distance threshold used to construct the building's adjacent spatial graph.

Table 5.9: Performance comparison between the xViews' first-place (XFP) solution and our proposed technique applied to our localization method.

Backbone	**Method**	**xView2 Score**	**Damage F1**	**Localization F1**
ResNet34	XFP	0.7913±2.6E-3	0.7602±3.7E-3	0.8640±2.7E-4
	Ours	**0.8007**	**0.7677**	**0.8777**
DualPathNet92	XFP	0.7941±5.6E-3	0.7645±7.8E-3	0.8630±1.0E-3
	Ours	**0.8024**	**0.7701**	**0.8777**
SENet-154	XFP	0.7913±6.3E-3	0.7604±9.2E-3	0.8635±6.1E-4
	Ours	**0.8082**	**0.7785**	**0.8777**
SEResNeXt-50 (32x4d)	XFP	0.8003±1.6E-3	0.7731±2.1E-3	0.8638±5.9E-4
	Ours	**0.8128**	**0.7849**	**0.8777**

Performance Comparison with Segmentation Methods: Table 5.9 summarizes our best results compared to the XFP team solution using the xView2 score metric. The XFP solution achieved first place in the xView2 competition using an ensemble approach, combining the predictive power of several trained architectures. Despite the significant differences in training approaches between our solution and XFP, our study aims to provide a fair comparison by training our localization method and comparing the performance of our proposed architecture to that of the single-trained damage assessment model, using the XFP team's pre-trained weights.

Considering that XFP trained each model under three different seeds, an average of the performance is provided, followed by the standard deviation.

Our solution trained on the single model SEResNeXt-50 (32x4d), together with our single model localization technique, achieves equivalent performance (i.e., 0.8128) to the ensemble method used by XFP (i.e., 0.8119). Our proposed approach enabled the predictive capacity of two trained models (localization and damage assessment) to match the performance of an ensemble technique composed of eight semantic segmentation models.

The classification maps illustrated in Figure 5.10 offer a visual comparison of XFP's solution using an ensemble method and our best-performing approach of equivalent performance. The comparison demonstrates how pixel-by-pixel segmentation methods produce salt and pepper noise in classification maps, making them difficult to interpret. Despite being applied to the predictions of a segmentation-based localization method, our instance-based approach puts a strong emphasis on identifying the damage to individual buildings. It produces predictions that are easier to interpret.

While the semantic segmentation approach can generate highly-performant pixel-level building localization and damage results, the assumption above might not always hold. We can observe in the results that multiple buildings are connected into one large region and that a building is split into multiple parts in the results. Meanwhile, it is a natural way to assess the building at the instance level in the real world. Therefore, we propose leveraging the instance segmentation model, which directly produces all the pixels of each building instance.

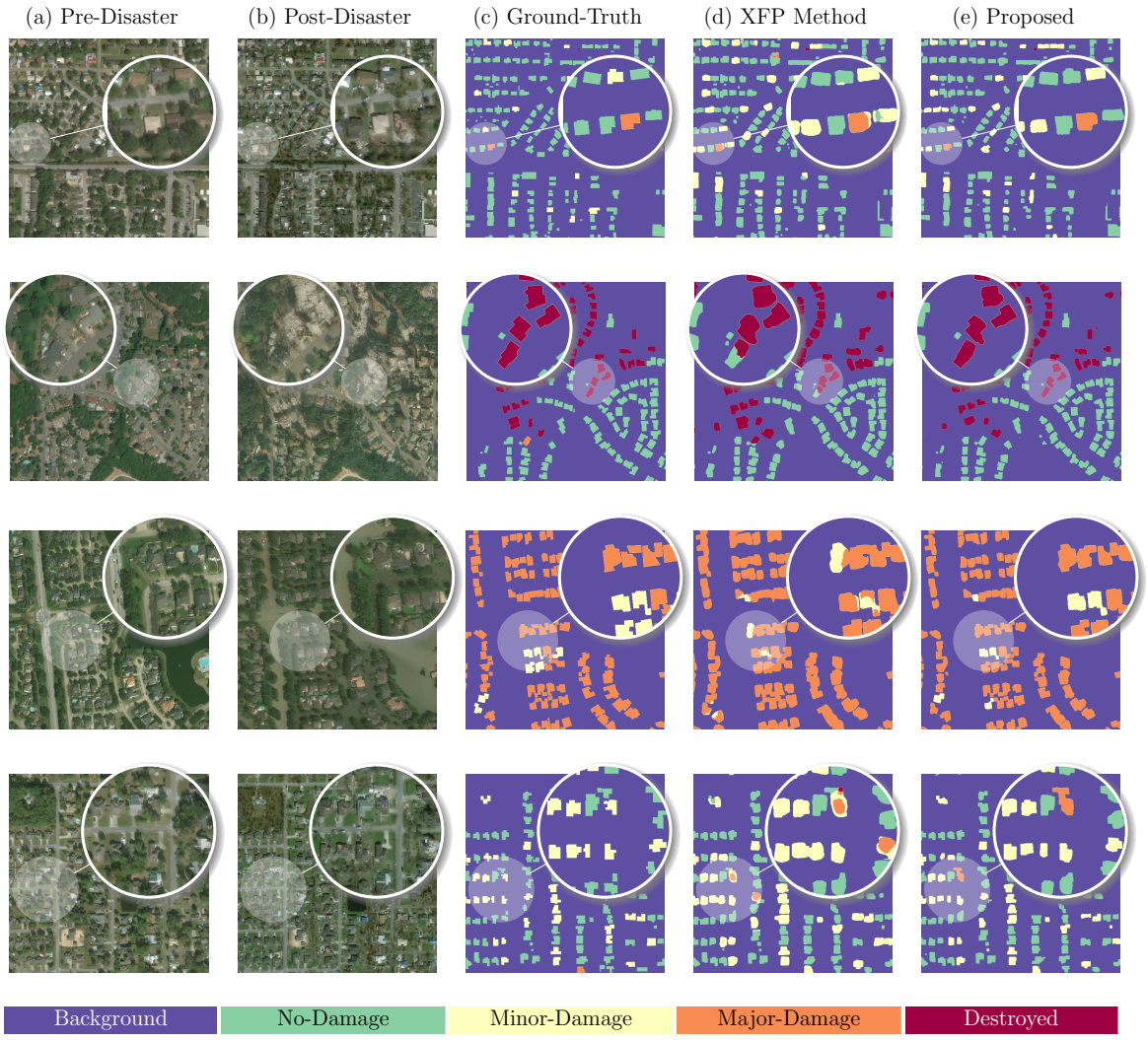


Figure 5.10: The pre- and post-disaster image tiles followed by ground-truth and classification maps produced by XFP's top solution and our proposed instance-based approach applied to our localization model's predictions. High-resolution color is suggested for the best viewing experience.

5.4 Conclusion

We demonstrate using pictures using a cutting-edge AI model trained to recognize 19 diagnostic morphological characteristics of honey bee wings. The model's performance shows that cutting-edge AI keypoint detection methods may be effectively used to study geometric morphometrics to identify honey bee species and subspecies.

The transfer learning approach was effectively used for both high and low-quality pictures in our current testing, allowing us to build correct models quickly. We also show that these keypoints are essential to help extract fine-grained characteristics that aid in subspecies classification.

Moreover, we demonstrate how we can identify and fuse fine-level patterns from the data at multiple levels of the deep learning architecture, addressing both local and global interactions of the data's characteristics to produce a more accurate prediction for samples that are difficult to tell apart. The application and fusion of disaster knowledge can provide improved situational awareness tools. In this analysis, by consolidating the satellite images taken before and after a disaster and finding the correspondence between the characteristics of the picture pair, a two-stream CNN network is leveraged to determine the degree of damage to residential homes. Using a curated dataset large-scale dataset, our model's performance and operational capacity are demonstrated.

CHAPTER 6

WEAKLY-SUPERVISED TRAINING

The speed at which the data is being collected far outpaces the rate of producing the well-curated expert label sets necessary to train a successful model using current methodologies. Recently developed tools and technologies capture more high-quality data at a rapid pace. It is possible to gather more data today than it was a few years ago, but the data collected is also quickly evolving in response to our changing environment, meaning historical data may rapidly become obsolete. For instance, new buildings and infrastructures are built daily to accommodate a growing population. People's interests and behavior also constantly change, following very distinct trends compared to a few decades ago. It is vital to develop supervised learning algorithms that effectively train datasets with noisy or limited labels.

This chapter proposes using a weakly-supervised model with the premise that the label set is restricted and may include instances of mislabels and mistakes. Random additive zero-centered Gaussian noise is injected throughout the training process to generate synthetic perturbations in the target label, supplement the data, and prevent over-fitting. To build a model that is resistant to the noise found in real-world data, such as coordinated position errors and crowd-sourced mislabeling or biasing, these small perturbations must be included in the training process. The component we present further combines a weakly-supervised deep learning technique with a new label propagation model to achieve its goals. The label propagation method enhances the training data by assigning labels to previously unlabeled data to improve the model's contextual awareness.

6.1 Weak-Supervision with Noise Regularization for Grid Targets

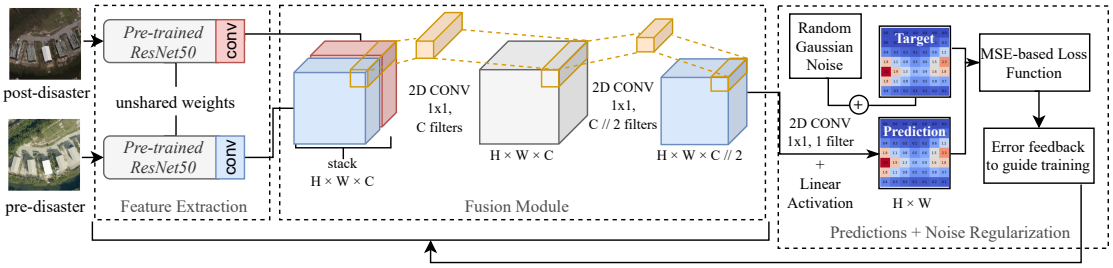


Figure 6.1: Proposed two-stream CNN architecture for the weakly-supervised damage assessment model applying the proposed fusion module to learn the deep feature correspondence at each feature location of the input image pair. Additive Gaussian noise is randomly applied to the target patches to help the model generalize better.

Damage assessment is a preliminary on-site survey of damages or failures caused by an accident or natural occurrences such as a hurricane, tsunami, or earthquake. These damage analyses are often conducted right after a disaster to document the degree of damage that can be replaced, repaired, or recovered. It can also estimate the time needed for repair, replacement, and rehabilitation. The first glimpse of the destruction caused by a disaster incident is made available by high-resolution satellite imagery or aerial photographs that allow experts to produce precise estimates of damage to the infrastructure without being physically present on-site.

Technological developments in remote sensing survey instruments [220, 221, 222], such as satellite images and aerial photographs, have enabled emergency responders to perform a comprehensive damage assessment quickly and remotely [223]. After a disaster event, government agencies such as Federal Emergency Management Agency (FEMA) conduct a preliminary building damage assessment through (1) predictive modeling to estimate probable damages; and (2) visual using the imagery captured

post-event to assess actual damages [108]. Nevertheless, manually identifying the impacted properties can be a slow and laborious job when disasters cover a wide land area.

Deep learning methods demonstrated great success in various research areas [8, 224, 1, 225, 209]. The Convolutional Neural Network (CNN) is a well-known architecture that has achieved tremendous breakthroughs in image recognition and has been the preferred method for developing damage assessment models using optical remote sensing data. However, recently proposed approaches that apply deep learning for building damage assessment are limited due to their reliance on the availability of accurate geometries illustrating the structure's footprint on the map. High-quality pre-disaster images are expensive to produce and may become outdated as new structures are developed.

We proposed to apply weak supervision in detecting the damaged building and the classification of the level of damage. The proposed work uniquely considers a scenario where a rough estimate of the damaged building's location and the degree of damage is the only data available to train the model. Such an approach will be valuable for rapidly identifying damaged buildings of interest and possibly expanding benchmark datasets for further studies without reviewing every structure in the image set.

An end-to-end convolutional neural network is developed in this work to automatically learn how to extract and fuse the characteristics from the images captured over an affected region before and after a disaster event. The assumption is to train a model under a scenario where data is minimal and geometric building footprints are scarcely available. Given a pair of images, the proposed model's objective is to generate a two-dimensional predictive patch in a regression-based approach. Each cell from the patch will contain a value of the predicted level of damage—the higher

the value, the greater the damage. The proposed approach allows the model to independently learn the specific patterns in the image that belongs to a building without the apriori knowledge of the building's footprint.

6.1.1 Motivation

Previously introduced building damaged scale classification research has been traditionally centered primarily on classifying damaged buildings into two categories (i.e., intact or destroyed) while also focusing on a single damage type due to a lack of well-curated benchmark datasets. Nonetheless, more recent research has explored a wider variety of damage levels, and datasets such as xBD [17] have allowed researchers to develop more complex models that can address the challenges unique to this direction.

Building damage assessment research is often divided into two categories: instance-level damage categorization [10, 194, 9] and pixel-level semantic segmentation [199, 200, 226]. Instance-level classification takes as input the image containing the overhead view of the building and aims to predict the level of damage the building sustained. The recommended method for pixel-level classification is semantic segmentation, which attempts to identify the building footprint and the degree of damage at each pixel level. The benefit of detecting damage at the pixel level is obtaining more fine-grained findings.

Building damage assessment techniques from previous works rely on the availability of high-quality geometry of the building footprint. However, the available building footprints may rapidly become outdated as new infrastructures are built while old ones are demolished. Pre-disaster images may become outdated, implying

localization models may not correctly identify the location of the newly developed buildings to make the correct assessment.

6.1.2 Data Pre-processing

The damaged point values are encoded in an ordinal format, starting with 1 as the lowest level of damage—the higher the value, the more extensive the damage. These values are multiplied by 100 and then placed in a grid on the spot where the damaged building is in the image to create the target label patch. A Gaussian kernel smooths out the point values in the grid, giving the damage more ground coverage of the surrounding area. Previously proposed works have demonstrated the region's importance surrounding the damaged structure when making predictions about the level of damage a building has sustained [9]. The Gaussian smoothing process is similar to the average filter but uses a particular kernel representing a Gaussian curve form.

The Gaussian smoothing function G formula is shown in equation 6.1.

$$G(d_x, d_y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{d_x^2 + d_y^2}{2\sigma^2}} \tag{6.1}$$

Where d_x is the horizontal axis distance from the origin, d_y is the vertical axis distance from the origin, and σ is the Gaussian distribution's standard deviation. During training, additive zero-centered Gaussian noise is randomly applied to the target label patch to create synthetic perturbations in the data and mitigate over-fitting. These small perturbations are meant to aid in training a robust model to the noise often found in real-life data [227] such as Global Positioning System (GPS) location errors.

6.1.3 Framework Configuration

Feature Extraction

As shown in Figure 6.1, our proposed end-to-end framework is configured as a two-stream CNN network that implements a fusion module to learn from the correspondence between each image's feature vector. Each stream uses the ResNet50 architecture [38], pre-trained on ImageNet, to extract features from the last convolutional layer and obtain a feature vector that is correspondent to parts of the 2-D image. Both networks' pre-trained weights are entirely fine-tuned to the new damage-assessment datasets. Moreover, the weights between the networks are unshared, as this has been demonstrated by previous work [10] to allow each stream the flexibility to fine-tune their weights accordingly and achieve better results individually.

Fusion Module

Following the feature extraction step, the fusion module is trained to learn the deep correspondence between both feature vectors using the capabilities provided by the one-by-one (1×1) 2-D convolution technique [185]. The two-stream networks in the feature extraction phase take as input an image pair and, from its last convolutional layer, generates the feature vectors $F^{pre}, F^{post} \in \mathbb{R}^{H \times W \times D}$, where H, W, D are the height, width and the total number of channels from the corresponding feature vector. The objective is to develop a fusion module, or function, such that $f : F^{pre}\ F^{post} \rightarrow P$, where $P \in \mathbb{R}^{H \times W}$ is the predictive output patch. The fusion model requires the vectors F^{pre} and F^{post} to first be stacked at the same spatial location (x, y) across channel d, namely:

$$F^{\text{stack}}_{x,y,2d} = F^{pre}_{x,y,d}, \qquad F^{\text{stack}}_{x,y,2d-1} = F^{post}_{x,y,d}, \tag{6.2}$$

where $\mathbf{F^{stack}} \in \mathbb{R}^{H \times W \times C}$ and $C = 2D$. It is also assumed $1 \leq x \leq H$, $1 \leq y \leq W$, $1 \leq d \leq D$. As shown in Figure 6.1, the 1×1 convolution that follows works as a coordinate-dependent transformation implemented to takes as input the stacked feature vectors and define the correspondence at each spatial location (x, y). This convolution approach leads to dimensionality reduction, with its combination being mathematically equivalent to a multi-layer perceptron at each (x, y) [185].

While the first 1×1 convolution layer takes care of learning the deep correspondence between F^{pre} and F^{post}, the 1×1 convolutional layer that follows further refines the fused features and reduces the dimensions. In contrast, the last convolutional layer applies a single 1×1 convolutional layer followed by a Linear activation that generates the final predictive patch P.

Predictive Results Post-processing

When running inference, the proposed model makes overlapping strides throughout the test dataset and outputs the predictive patches for each stride. These predictive patches are then merged, with the maximum value of the overlapping cells calculated to generate the final predictive heatmap, where the higher the values of the output grid cell, the higher the damage sustained by the building located in that specific area.

Table 6.1: Data summary of the Irma data and xBD data.

No.	Concepts	Irma	xBD
		# of damage instances	
1	Affected	1219	-
2	Minor Damage	2739	36860
3	Major Damage	1082	29904
4	Destroyed	577	31560

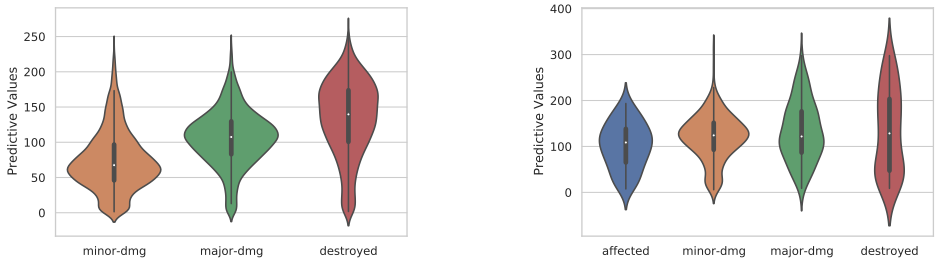
6.1.4 Experimental Analysis

Experimental Setting

In this dissertation, two datasets are tested on the proposed methods. The Irma data samples are split into three non-overlapping areas for training, validation, and testing. Two of the lower and middle keys, Sugarloaf Key and Cudjoe Key, are selected as validation and testing set areas. The xBD satellite images are also split in a similar manner following the original xBD data split provided in train, test, and holdout splits in the 80/10/10% split ratio. In the proposed approach, the test set is used as a validation set to make an unbiased evaluation of our model while training and storing the best-performing results. The holdout set is used to test the final already trained model.

Ablation Study

In this paper, two datasets are tested on the proposed methods. The Irma data samples are split into three non-overlapping areas for training, validation, and testing. Two of the lower and middle keys, Sugarloaf Key and Cudjoe Key, are selected as validation and testing set areas. The xBD satellite images are also split similarly to the original xBD data split provided in train, test, and holdout splits in the 80/10/10% split ratio. In the proposed approach, the test set is used as a validation set to make an unbiased evaluation of our model while training and storing the



(a) xBD data (b) Irma data

Figure 6.2: Violin plots of the sampled damaged points from predictions made on the xBD (left) and Irma (right) test split, grouped by the different levels of damage.

best-performing results. The holdout set is used to test the final already trained model.

Using the Mean Squared Error (MSE) formula, two loss functions are defined and assessed for the proposed approach—the patch loss $\mathcal{L}_{patch}$ function and the pixel loss $\mathcal{L}_{pixel}$ function. As demonstrated in equation 6.3, $\mathcal{L}_{patch}$ first calculates the losses for each grid cell (i, j) at each patch level P_m, where $(i, j) \in P_m \in \mathbb{R}^{N \times 2}$ and N is the total number of cells in the grid. These losses are then summed up across a batch and divided by the total number of patches M in the batch to obtain the average loss from each patch level.

$$\mathcal{L}_{patch} = \frac{1}{M} \sum_{m=0}^{M} [\frac{\sum_{i,j \in P_m} (Y(i,j) - \hat{Y}(i,j))^2}{N}] \tag{6.3}$$

As an alternative, the pixel loss $\mathcal{L}_{pixel}$ function demonstrated in equation 6.4 calculates losses for the individual cells in the grid, treating each predicted cell as its individual sample. The variable K represents the total number of individual cells in each training batch and is also assumed to be equal to $M \times N$.

$$\mathcal{L}_{pixel} = \frac{1}{K} \sum_{k=0}^{K} (Y_k - \hat{Y}_k)^2 \tag{6.4}$$

The performance improvements made by the proposed fusion module are also corroborated by comparing the results from the post-only model and pre-post fusion model configurations detailed as follows:

Post-only configuration: A single CNN model, based on the ResNet50 architecture and pre-trained on ImageNet, is entirely fine-tuned to extract the features from the images taken after the disaster event. The original classification head is removed and replaced with a 1×1 2-D convolutional layer followed by a linear activation function to output the predictive patch.

Pre-post fusion configuration: This configuration makes use of the proposed fusion module highlighted in Figure 6.1, more details about this configuration have to be found in Section 6.1.3.

Each epoch, the model is trained on the entire training set using a small batch size of 32 samples to regularize further and improve the model's generalization capability [196]. As the model trains, it is evaluated at the end of each epoch on the validation set with the best performing model with the lowest validation loss saved—models are trained for no longer than 100 epochs. In these experiments, the model's weights are optimized using the Stochastic Gradient Descent with an initial learning rate of $\eta = 0.001$. During training, the learning rate is multiplied by a factor of 0.1 after no improvements to the validation loss for 10 consecutive epochs.

Results and Discussion

For testing purposes, values for damaged location points in the testing set are sampled using bilinear interpolation [228] from the final predictive heatmaps' output predictive value cells to compare with the labeled assessment data. In bilinear interpolation, given the sampling point's location, the four cell centers from the predictive input patch that is nearest to the sampled cell's center are weighted based

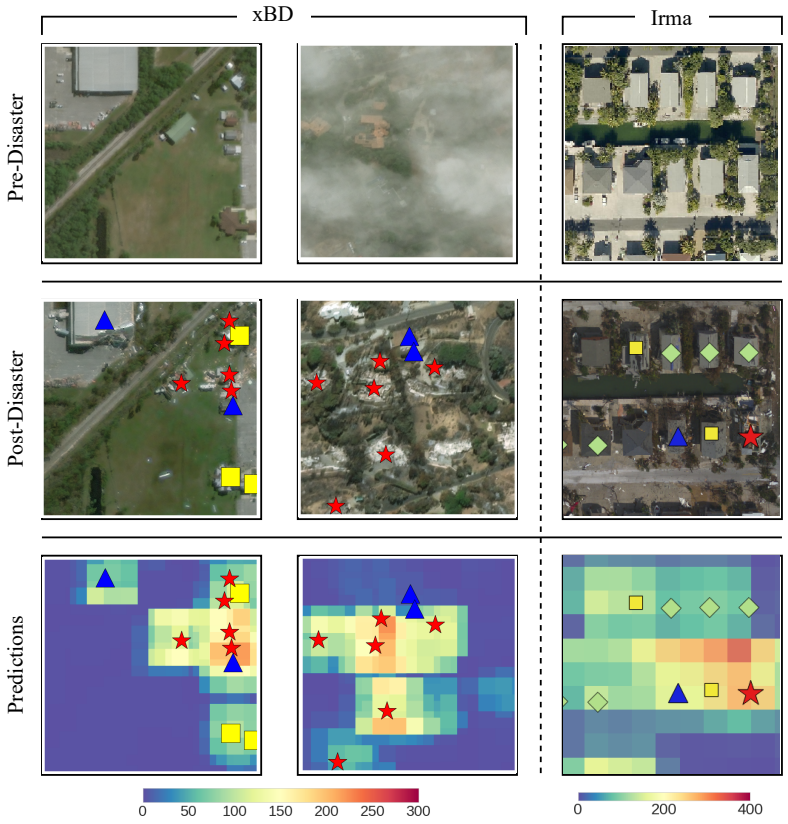


Figure 6.3: Qualitative results summary of the model's output predictive values on the test set for the xBD and the Irma data. The point locations for the damaged buildings are overlaid on the post-disaster images and the model's predictive patch. The legend for the damage point labels is as follows: ◊ - affected, □ - minor-damaged, △ - major-damaged, and ⋆ - destroyed.

on the distances and then averaged. Table 6.2 and Table 6.3 summarize the performance results for the xBD data and the Irma data test splits using well-known regression metrics MSE and Mean Absolute Error (MAE). These metrics are utilized to compare the performance among the model configurations described in previous sections. The closer the predictions made by the model are to the target damage instances, the more confidence can be placed into the model to more accurately detect the different damaged buildings.

The pre-post fusion model has consistently achieved the best performance. Moreover, losses calculated from each pixel level may perform better on smaller datasets due to MSE's sensitivity to outliers, which is essential when detecting the minority classes. Noise regularization plays an essential role in augmenting the training data and helping to train a more robust model. As part of the ablation study, the proposed pre-post fusion model is tested with and without noise regularization, demonstrating the performance improvements achieved by applying it. Although the proposed noise regularization hinders the performance of the smaller and more limited Irma data, it has proven to be an essential technique when working with larger datasets such as xBD.

Furthermore, the proposed patch-level loss function has also helped the model achieve better performance and reduce errors. Unlike the pixel-level loss function, the patch-level loss is not as sensitive to noise and places more weight on the surrounding regions.

Table 6.2: Performance summary on the xBD data.

Method	**Loss Function**	**Noise Reg.**	**MSE**	**MAE**
post-only	$\mathcal{L}_{pixel}$	✓	1.2259	0.9951
pre-post fusion			1.1867	0.9900
		✓	1.1285	0.9344
post-only	$\mathcal{L}_{patch}$	✓	1.2111	0.9828
pre-post fusion			1.1437	0.9436
		✓	**1.1282**	**0.9266**

Figure 6.2 demonstrates the violin plots of the best performing model configurations for the xBD and Irma data—its broader segments represent members of the population that are more inclined to take on the given value; the skinnier sections

Table 6.3: Performance summary on the Irma data.

Method	**Loss Function**	**Noise Reg.**	**MSE**	**MAE**
post-only	$\mathcal{L}_{pixel}$	✓	1.7227	1.4860
pre-post fusion			1.4714	1.2342
		✓	1.5663	1.3203
post-only	$\mathcal{L}_{patch}$	✓	1.5356	1.2722
pre-post fusion			**1.3893**	**1.1235**
		✓	1.5282	1.3039

express a lower likelihood. It can be observed that the weakly-supervised model can find an evident pattern among different damage levels —this is especially true with the results of the xBD data. Even if there is an offset between the target values and the predicted values, it is clear that the majority of the sampled predicted points are grouped on a segment following an ordinal paradigm—the higher the value from the predictive cell, the higher the level of damage sustained by the building located in that cell.

The model has successfully, and through implicit means, learned to detect different damage levels sustained by buildings. Figure 6.3 illustrates the qualitative results generated by the proposed model on the xBD and Irma datasets. These predictive outputs serve as the interpretable visualizations that can be leveraged for rapid labeling jobs and further studies in the weakly-supervised building damage assessment effort. Further refinement must be done to the predictive outputs to train a model that can further separate the individual buildings and detect the damages they sustained.

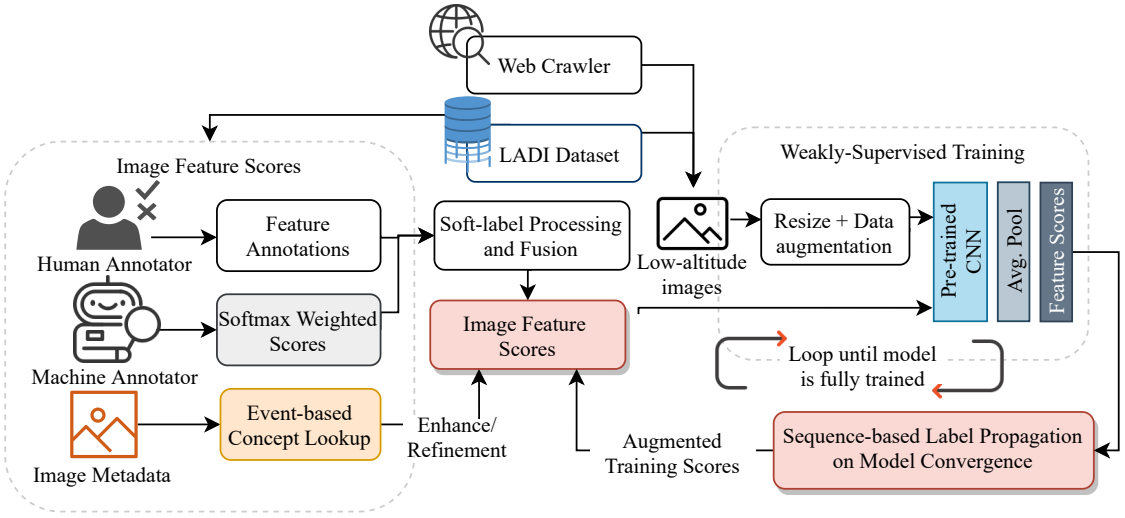


Figure 6.4: The proposed weakly-supervised deep learning framework implements a feature fusion and automatic propagation of feature scores based on the spatio-temporal information obtained from the low-altitude image's metadata.

6.2 Weak-Supervision with Label Propagation

This dissertation proposes a weakly-supervised approach to training a deep neural network on low-altitude imagery with highly imbalanced and noisy crowd-sourced labels. We further use the rich spatio-temporal data obtained from the pictures and its sequence information to enhance the model's performance during training via label propagation. Our approach achieves the highest score among all the submitted runs in the TRECVID2020 Disaster Scene Description and Indexing (DSDI) Challenge, indicating its superior capabilities in retrieving disaster-related video clips compared to other proposed methods.

Pictures or videos captured from a low-altitude aircraft or an unmanned aerial vehicle are a fast and cost-effective way to survey the affected scene to assess a catastrophic event's impacts and damages quickly. Using advanced techniques like deep learning, now more than ever, it is possible to automate the disaster scene description and identify the features in captured pictures or recorded videos to bring

situational awareness. However, building a large-scale, high-quality dataset with annotated disaster-related features for supervised model training is time-consuming and costly. This paper proposes a weakly-supervised approach trained on highly imbalanced and noisy crowd-sourced labels to advance the curation and retrieval of the low-altitude imagery. We further use the rich spatio-temporal data obtained from the pictures and its sequence information to enhance the model's performance during training via label propagation. Our approach achieves the highest score among all the submitted runs in the TRECVID2020 Disaster Scene Description and Indexing (DSDI) Challenge, indicating its superior capabilities compared to other proposed methods.

The advent of deep learning tools and techniques, especially the Convolutional Neural Network (CNN) [229, 1], has revolutionized image and video recognition and greatly improved object detection accuracy and robustness. Considering how images and videos are a prevalent way for emergency responders to survey affected areas after a natural disaster quickly, it is no surprise that deep learning methods such as CNNs are being applied to automate the curation and retrieval of the captured images.

Although research on automatic disaster scene descriptions from images has become more prevalent in recent years, most existing approaches are confined to one disaster type. In addition, they are often incomparable due to the lack of well-curated datasets and benchmarks. This has recently changed with the introduction of large-scale disaster datasets such as the Incidents Dataset [168] and LADI [18]. The Incidents Dataset is well-curated; however, it focuses on a ground-level perspective, which does not assist with the significant intra-class variances shown on low-altitude images. Most of the previously proposed methods in the disaster scene description from both image and video data also focused on the ground-level point

of view. These methods often aim to address challenges commonly found in the disaster image data by developing sophisticated models such as adversarial data augmentation to deal with the limited data [230] or standard techniques that put a higher penalty to errors on the minority class to address class-imbalance issues [231, 209].

The LADI dataset features a wide range of low-altitude images. Some of these images have noisy annotations with many different samples per class. Moreover, some objects and features in LADI are shown at different sizes and angles depending on the altitude at which the picture was taken, making them hard to identify. The earlier research on the disaster scene description from low-altitude images tested different supervised methods by considering the image's optical properties [232, 233, 179]. More recent studies started to explore an ensemble learning approach to tackle the class imbalance and noisy-label issues [171, 172, 234].

Moreover, the previously proposed methods seldom leverage the rich spatio-temporal information from data. They have yet to exploit the sequential-based information of the low-altitude images to improve the model performance during training. Our proposed framework leverages the weakly-supervised deep learning approach with a unique label propagation model that enhances the training data as the model learns and uses the spatio-temporal information to improve the contextual awareness of the model.

6.2.1 Motivation

A catastrophic event or accident may have disastrous implications, such as making some of the most impacted regions utterly inaccessible due to outages in communication lines and disruptions to street-network infrastructures. Furthermore, the

availability of trustworthy and accurate information is a crucial challenge for emergency management. However, depending on the scale of the disaster, the large volume of collected data and the limited time under a disaster scenario make it extremely challenging to identify regions that should be prioritized rapidly.

Civil Air Patrol (CAP) supports U.S. communities going through an emergency response by taking pictures or recording videos from a low-altitude aircraft as a crucial and inexpensive method for the Federal Emergency Management Agency (FEMA) to quickly and effectively obtain the imagery necessary to survey the affected region. The footage is often captured from military aircraft, primarily cargo aircraft, tankers, or helicopters [235], and more recently, drones as well [236].

Given the massive volume of data being gathered, it is becoming more vital to develop sophisticated tools and systems to curate all the information [237, 238]. Several large-scale disaster datasets, including the Incidents Dataset [168], LADI (Low Altitude Disaster Imagery) [18], etc., have been recently released to stimulate the development of new research and technologies in this field. However, the current public benchmarks suffer from the data scarcity problem, and further analysis methods are required to tackle it.

6.2.2 Proposed Framework

In this dissertation, a weakly-supervised learning framework for disaster scene description as illustrated in Figure 6.4 is proposed to address the challenges that data labels are limited in both quantity and quality. In light of these limitations, classifiers pre-trained on other well-curated benchmarks are leveraged to supply supervision signals by connecting their predicted concepts to the target features at the semantic level. Soft-labels are created initially from human workers' annotations. The

more workers who annotate an image with a target feature, the greater the weight allocated to the image under that target feature. These soft-label features are then fused with SoftMax weights of well-known deep-learning methods that have been pre-trained with well-curated large image datasets. The final soft-labels defining the likelihood of an image possessing a given feature are propagated throughout the training dataset to enrich the training data appropriately. While the deep learning-based model learns, it helps identify more samples of a specific feature and expand the label set. The proposed method reduces the difficulty of obtaining well-curated expert hand-labeled data sets, which may be expensive or unfeasible. Instead, low-cost weak labels are used to leverage them to develop a robust prediction model, even if they are wrong. The proposed approach estimates the likelihood of a particular disaster or environment-related feature being present inside a low-altitude image or video.

Feature Score Engineering

Worker Annotations: The LADI dataset uses a hierarchical labeling approach featuring five general categories, including *damage*, *environment*, *infrastructure*, *water*, and *vehicle*. Within each category, features of more specific categories are annotated. Using the Amazon Mechanical Turk (MTurk) service [174], a subset of the LADI dataset, representing more than forty thousand images, was hand-annotated by human annotators.

Assuming that the data is either labeled by non-expert human workers through crowd-sourcing or obtained from a web crawler, label engineering is a critical initial step in reducing label noise and preventing erroneous labels from deceiving the model. Given an image i in the dataset, it may be labeled by one or more workers as containing a feature `f`. However, not all the worker's labels can be treated with an

equal level of confidence. Let $C_{i,\mathtt{f}}$ be the number of workers who labeled the image i as containing feature $\mathtt{f}$. Each image's feature score is $S_{i,\mathtt{f}} = (C_{i,\mathtt{f}} - C_{\mathtt{f}}^{\mathtt{min}})/(C_{\mathtt{f}}^{\mathtt{max}} - C_{\mathtt{f}}^{\mathtt{min}})$, where $C_{\mathtt{f}}^{\mathtt{min}}$ and $C_{\mathtt{f}}^{\mathtt{max}}$ are the minimum and maximum counts of workers for all the annotated images with feature $\mathtt{f}$. The soft-score function is formulated under the assumption that there will be at least one human worker annotating an image for target features under the same category. The assumption is that $C_{\mathtt{f}}^{\mathtt{max}} > 0$. Under this assumption, $C_{\mathtt{f}}^{\mathtt{max}} = C_{\mathtt{f}}^{\mathtt{min}}$ implies that all the positive samples are annotated by the same amount of workers. Thus, we assign the labels of all these samples as 1, i.e., $\forall i, S_{i,f} = 1$.

In our investigations, we employ these normalized soft-label vectors as the ground truth confidence. Soft-labels provide a model with more information about the relevance of each target feature. Such a strategy works well in a ranking problem scenario, provided that the goal is to help index the most relevant images. However, these crowd-sourced human labels are highly imbalanced. Because of the extreme disparity between different labeled samples, the calculated $S_{i,\mathtt{f}}$ may be imprecise for significantly under-represented features. In addition, some images also bear incorrect or ambiguous labels. Hence, the dataset requires further enhancements by adding new data and new information.

Machine Annotations: The LADI dataset includes several machine-generated feature scores from commercial and open-source image recognition platforms to provide additional knowledge for various features found in the images. These feature scores are in the form of SoftMax weights that can be defined as follows.

$$\sigma(\hat{x})_i = \frac{\exp(x_i)}{\sum_{k=1}^{K} \exp(x_k)} \tag{6.5}$$

Where given the input vector $\hat{x} = (x_1, ..., x_K) \in \mathbb{R}^K$, the equation applies a normalization term to output probability distribution for K classes. The SoftMax weights

are thus numerical scores indicating the relative confidence of the pre-trained model in detecting the existence of a certain feature, and these machine annotations contain the names of the detected features, allowing us to match them with the features in LADI via semantic similarity.

The first machine annotator is a ResNet50 model [38] pre-trained on Places365 [175], which includes 365 categories of ordinary scenes. It is crucial to improve the efficacy of detecting the scenes and environments in the LADI imagery. Many features present in LADI are broad terms and can be matched with many available features in Places365 according to the semantic meaning. For instance, the concept for "building" has a close semantic meaning with the Places365 concepts, including *apartment building outdoor*, *basement*, *beach house*, *building facade*, *construction site*, *downtown*, *residential neighborhood*, *roof garden*, *skyscraper*, etc. Therefore, many matched concepts could be safely regarded as "building" for disaster scene description.

Machine-generated annotations from Google Cloud Vision (GCV) [239] are also part of the machine annotators. The GCV API provides robust pre-trained machine learning models for instantly assigning labels to images and classifying them into millions of preset categories. The GCV *label detection* and *web entity detection* services scores are available for a subset of the LADI dataset.

We further used the predictions from the YOLOv4 [240] model pre-trained on the COCO dataset [241]. The annotations supplied by the YOLOv4 model trained on COCO contain significant characteristics like car and truck and have shown to be significant in improving the *vehicle* category model.

Information from other sources was retrieved for a subset of the highly underrepresented features. Crawling for more data helps alleviate some of the training dataset's imbalanced problems and mistakes found within the labels. A small num-

Table 6.4: A summary of the conflicts between LADI's initial labels specified by human workers and the labels rectified utilizing external databases, namely FEMA and NOAA CDO.

Feature	**P→P**	**P→N**	**U→P**
(1) Damage (misc)	5146 (12.7%)	7839 (19.3%)	-
(2) Flood/water	12554 (30.9%)	4236 (10.4%)	-
(6) Smoke/fire	4 (0%)	1614 (4.0%)	100 (0.2%)
(12) Snow/ice	0 (0%)	115 (0.3%)	198 (0.5%)

ber of sample images under these features are crawled using an image search engine, such as Microsoft Bing Image[1], while also making sure the queries contain words such as *drone* and *aerial* along with the target feature name. The noise from the crawled images is reduced by applying the CNN model trained on the human workers' limited labeled data and selecting the images that are most relevant to their target feature (i.e., score > 0.5). The scores from all relevant crawled images are then set to 1.

Metadata Concept Lookup: To include more concepts relevant to real-life events, we further utilize time and location metadata obtained from each image. The focal length (F), altitude (A), latitude, longitude, and camera type are all contained in the data that may be retrieved from an image's metadata, provided that the image's format supports the Exchangeable image file format (Exif). This information is useful to approximate the geographical region covered by the image taken from an airplane. Although there is no direct access via Exif to the height `H` and width `W` of the camera sensor, the camera model provided by the metadata was used to acquire this information from other sources. The simple trigonometry can specify the current footprint through the computation of width $= (A \times \texttt{W})/F$ and height $= (A \times \texttt{H})/F$ of the geographic region. The measured area is only a rough approximation of the area photographed, as illustrated in Figure 3.4, and is limited

[1]Microsoft Bing Image: `https://www.bing.com/images`

by its assumption that the camera takes the picture while being pointed directly downwards. The angle from which the image was taken is a required parameter to determine the exact geographical boundaries covered by the image. Nonetheless, several drones on the market today feature more detailed information regarding location captured and camera angle.

The image's time and location metadata offer multiple useful indications concerning the image's contextual content. Additional information about the photographed region can be accessed from open datasets by considering the particular incidents, locations, and special weather conditions that may have been recorded at the time and place the picture was taken. Open databases used to retrieve the images' contextual data are summarized as follows.

- **OpenStreetMaps OSM**: The computed geographical region represented as a polygon is used to index open-source geo-databases OpenStreetMaps [242] which provide a valuable location-based context of the aerial images captured. OSM tags are crowd-sourced and describe specific features of map elements. The more area covered by the image, the more tags it contains in terms of buildings, roads, etc. Bringing in the OSM data starts with collecting the total number of tags that fall inside the image's capture region and using a min-max normalization to provide more confidence over the images containing a higher number of a particular tag. Nonetheless, as illustrated in Figure 3.4, there is a noticeable shift between the computed region and the actual area shown in the image. Despite the limitations of the OSM's geographic information-based approach, our proposed methods of combining several modalities help generate more reliable scores. In addition, Figure 6.5 demonstrates the logistic regression fitted on the relationship between the LADI soft-labels and the matching OSM scores, further supporting our hypothesis that there is a

positive relationship between the relevant OSM tags and the target features that can be found in an image. The plot illustrates the most relevant target features that are semantically similar to the OSM tags. Section 6.2.2 details the matching procedure and aggregation method for the OSM scores.

- **FEMA**: FEMA Disaster Declarations are an excellent resource for records of historical disasters in the United States, such as coastal storms, earthquakes, fires, floods, hurricanes, tornadoes, and volcanic activities. The FEMA data is used to confirm that the images in the LADI dataset annotated with the specific reported damage correspond to the actual real-life incident. If an annotator or pre-trained model declares that a label contains specific damage caused by a type of disaster in contradiction with the FEMA records, the score for that feature is set to 0. Otherwise, the score's confidence is increased to 1. Table 6.4 summarizes the concordance and conflicts between the original worker labels and the labels rectified using the external databases, including FEMA. The positive-to-positive (P→P) category agrees that an image under a relevant target feature was annotated as positive. On the other hand, the conflicts between actual and rectified labels are shown through the positive-to-negative (P→N) category. The actual label was positive in conflict with the rectified negative label. In the unknown-to-positive (U→P) category, because of the low reliability of the original annotations, positive samples are drawn from the LADI unlabeled-set for the target features such as smoke/fire by using the information from the external databases. Later in Section 6.2.3, we show how a reliability measure further helps to explain the high degree of conflicts observed in the LADI label set for many target features, including smoke/fire and snow/ice.

- **NOAA Climate Data Online (CDO)**: NOAA allows for public access to the National Climatic Data Center's (NCDC) [243] data which is an archive of global historical meteorological and climatic data. For features related to the climate, such as snow, weather details were obtained from the NCDC API using the time and location information. Similar to FEMA, the NCDC data confirms that the annotation given to the image does not contradict the real-life event. For example, to ensure that a given image's snow feature is valid, the image's time and location are compared to the NCDC data. The feature score is adjusted to 1 if snow has been reported, or 0 otherwise. As shown in Table 6.4, similar to the smoke/fire target feature, snow/ice was further enhanced through the addition of positive samples from the unlabeled-set.

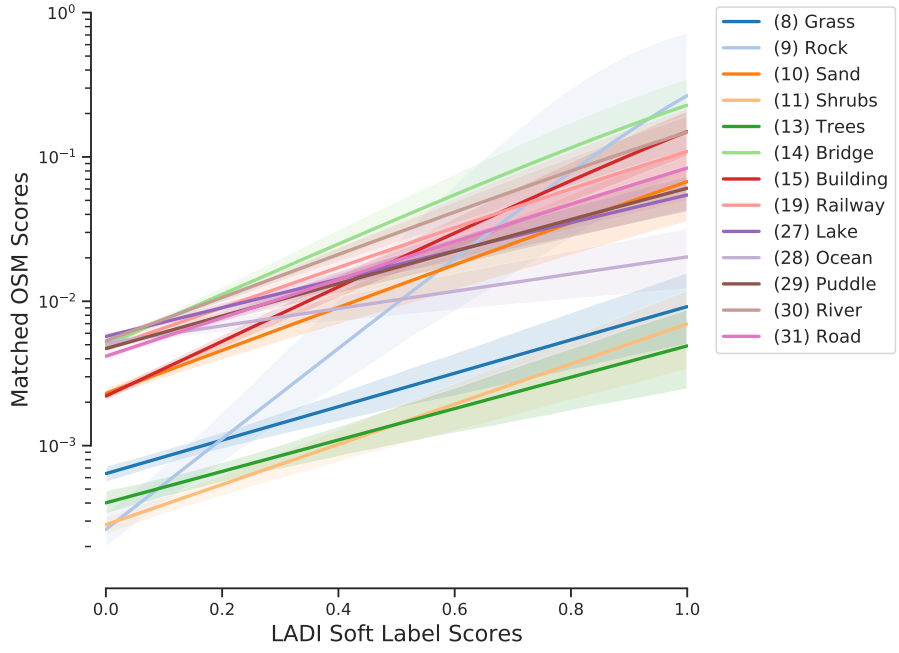


Figure 6.5: Logistic regression plot and a 95% confidence interval of the relationship between the LADI soft-labels and the matching OSM scores. Only the target features found to be semantically similar to the OSM tags are included in this plot.

Final Score Fusion

The model is trained to identify a particular feature inside an image and its confidence level by employing soft-labels. This also makes it easier to integrate soft-labels generated from human annotations with SoftMax weights supplied by various pre-trained classifiers assessed on the LADI dataset. Prior to the actual fusion of the scores, a semantic match of the feature's name in both LADI and the pre-trained model's own feature list is formed. Semantic similarity borrows techniques from Natural Language Processing (NLP), such as word embedding, to determine how similar two words are, even when they are not exact matches. The feature's word vectors were first generated using the pre-trained model from spaCy [244] followed by a pairwise computation of the vector's cosine similarity [245]. Since some features are composed of two or more words, spaCy helps to identify the unique words (or tokens) within the feature name and generate the word vector. Considering that the features found in the LADI dataset are very broad in concepts, we take advantage of the variety provided by the scores generated from both the pre-trained models and OSM tags. Selecting the relevant OSM tags is also done through semantic nearness by finding the similarity between two word vectors for the OSM tags and the target feature name in the vector space.

Because the SoftMax weights are a probability distribution that awards the highest score to the best-detected classification in each image, a min-max normalization is applied before working on the final score fusion. Let $\hat{w}_i^{\mathtt{f}}$ represent the word vector of the i-th word in the name that describes the target feature $\mathtt{f}$ and $\hat{w}_j^{\mathtt{p}}$ represents the word vector of the j-th word in the name that describes the auxiliary feature $\mathtt{p}$ in either the pre-trained classifiers or the OSM tags. All the word vectors are generated by spaCy [244]. The final score fusion is decided based on the distance of

the word vectors, as shown below.

$$S^*_{\mathtt{f}} = \max(S'_{\mathtt{f}}, \max_{\mathtt{p} \in \mathbb{P}_f} S_{\mathtt{p}}) \tag{6.6}$$

$$\mathbb{P}_f = \left\{ p \,\middle|\, \max_{i \in [1, N_{\mathtt{f}}], j \in [1, N_{\mathtt{p}}]} \frac{\hat{w}^{\mathtt{f}}_i \cdot \hat{w}^{\mathtt{p}}_j}{\|\hat{w}^{\mathtt{f}}_i\| \|\hat{w}^{\mathtt{p}}_j\|} > \vartheta \right\} \tag{6.7}$$

where $S^*_{\mathtt{f}}$ is the final score assigned to a certain image for feature $\mathtt{f}$, $S'_{\mathtt{f}} \in [0, 1]$ is the score of the image for target feature $\mathtt{f}$ integrating the human annotations and machine annotations, $S_{\mathtt{p}} \in [0, 1]$ is the score integrating pre-trained classifiers and OSM tags for the auxiliary feature $\mathtt{p}$, and $N_{\mathtt{f}}$ and $N_{\mathtt{p}}$ are the numbers of words that describe the target feature $\mathtt{f}$ and auxiliary feature $\mathtt{p}$, respectively. The final score fusion first checks that the cosine distance between $\hat{w}^{\mathtt{f}}_i$ and $\hat{w}^{\mathtt{p}}_j$ must be greater than a given threshold ϑ (0.5 in this study) for the concepts to be considered semantically similar. The final score for the target feature $\mathtt{f}$, i.e., $S^*_{\mathtt{f}}$, is thus the largest score among the original score $S'_{\mathtt{f}}$ and the scores of any auxiliary features whose names contain at least a word semantically similar to any word in the name of target feature $\mathtt{f}$, as illustrated in Equations (6.6) and (6.7).

Weakly-Supervised Training

Weak supervision discussed in this paper is a strategy that learns from the partially annotated and noisy labels and the low-quality information from various data sources. Our proposed system combines a novel label propagation approach with a weakly-supervised deep learning framework to improve the data quality as the deep learning model trains. The suggested technique aims to make acquiring well-curated expert hand-labeled data sets easier by using low-cost weak labels. Algorithm 1 illustrates the steps to train one of the categorical models in a weakly-supervised approach via label propagation. The proposed approach extracts deep features from the final convolutional layer. It then outputs a feature vector corresponding to

Algorithm 1 Weakly-Supervised Model Training implementing Label Propagation

1: $M_0^1 \leftarrow M_S, r \leftarrow 100, \sigma \leftarrow \infty, \rho \leftarrow 5, i \leftarrow 1$
2: **repeat**
3: $\quad \rho_i \leftarrow 0$
4: $\quad$ **while** $\rho_i < \rho$ **do** ▷ Train until convergence
5: $\quad\quad M^* \leftarrow Train(M_{i-1}^1; X_t, Y_t)$
6: $\quad\quad \rho_i \leftarrow \rho_i + 1$
7: $\quad\quad \sigma_i \leftarrow Loss(M^*; X_v, Y_v)$ ▷ Using Equation 6.8
8: $\quad\quad$ **if** $\sigma_i < \sigma$ **then** ▷ Keep the best model
9: $\quad\quad\quad M_i^1 \leftarrow M^*, \sigma \leftarrow \sigma_i, \rho_i \leftarrow 0$
10: $\quad\quad$ **end if**
11: $\quad$ **end while**
12: $\quad$ **for** each feature $\mathtt{f}_j \in F$ **do**
13: $\quad\quad \hat{j} \leftarrow TopScores(y_{i,j}, r)$ ▷ Fetch the top samples
14: $\quad\quad X_{\hat{U}} \leftarrow NearestSamples(X, \hat{j})$
15: $\quad$ **end for**
16: $\quad Y_{\hat{U}} \leftarrow Predict(M_i^1; X_{\hat{U}})$ ▷ Propagate the scores
17: $\quad X_t \leftarrow Merge(X_t, X_{\hat{U}}), Y_t \leftarrow Merge(Y_t, Y_{\hat{U}})$
18: $\quad i \leftarrow i + 1$
19: **until** *Stop Condition* is met ▷ Model is fully-trained

the 2-D picture using the InceptionV3 architecture, pre-trained using the ImageNet weights. The pre-trained weights of the networks have been completely fine-tuned to the new low-altitude image dataset. The model's original classification head is replaced with a dense layer followed by a sigmoid activation function to enable multi-feature score prediction indicating the likelihood that an image includes a certain feature.

The training process starts with the model M_0^1 initialized to the ImageNet weights M_S. From line 4, the model trains on the training dataset composed of the images X_t and scores Y_t, where X is the entire set of low-altitude images, whether labeled or unlabeled and $X_t \subseteq X$. In each epoch, the total loss $\mathcal{L}$ is calculated by aggregating the binary cross-entropy (BCE) loss across all the individual features

as follows.

$$\mathcal{L}(p_{\mathtt{f}}, q_{\mathtt{f}}) = -\frac{1}{\mid F \mid} \sum_{\mathtt{f} \in F} (p_{\mathtt{f}} \log(q_{\mathtt{f}}) + (1 - p_{\mathtt{f}}) \log(1 - q_{\mathtt{f}})) \tag{6.8}$$

where $p_{\mathtt{f}}$ is the probability (or soft-label) of the image containing the target feature `f`, and $q_{\mathtt{f}}$ is the predicted probability of the image containing `f` as calculated by the model. As the model trains, it is validated on the validation samples X_v in which its predicted values are compared to the target scores Y_v using the loss function. The variable σ keeps track of the lowest validation loss such that only the best model is kept at the end of the training process.

Once the model reaches a point where it is no longer improving for ρ consecutive epochs, the algorithm starts the label propagation process at line 12 by first sampling the top scores from each feature in X_t to later acquire the nearest samples. For r unique observations identified from X_t under a specific feature `f`, the algorithm finds the nearest unclassified samples and stores them in $X_{\hat{U}}$. Scores are propagated throughout the identified image's neighboring images. The idea is that if a picture taken at a particular moment contains a specific feature, the picture taken before and after it will most likely have the same feature. The training process is terminated when the *Stop Condition* is met, i.e., the model is considered to be fully trained. The stop condition is defined as the model not being improved after two consecutive label propagations.

6.2.3 Experimental Analysis

Dataset

This paper uses the LADI dataset, which consists of pictures captured from a low-flying aircraft by CAP and hosted by FEMA. The National Institute of Standards

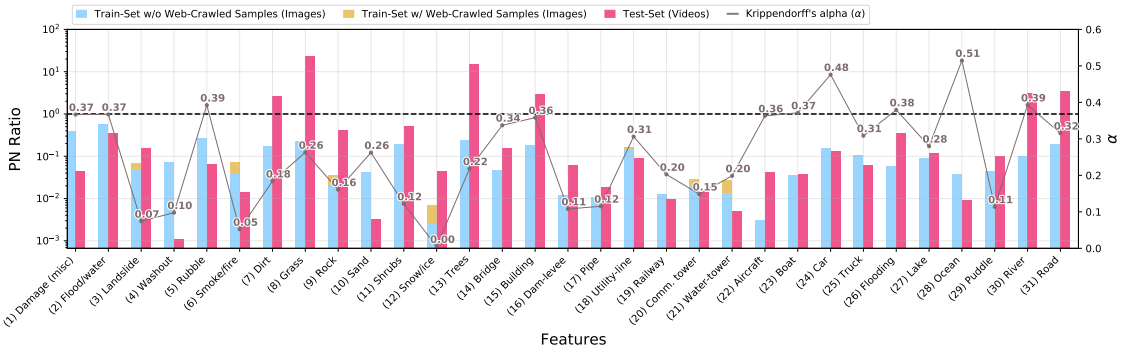


Figure 6.6: The PN ratios of the 31 target features from the LADI training and testing datasets before any of the proposed rectification techniques were applied. The reliability coefficient Krippendorff's alpha (α) indicates the measure of the agreement among the workers when annotating the training dataset for each target feature.

and Technology (NIST)'s TREC Video Retrieval Evaluation (TRECVID) competition released the dataset to participants in the middle of 2020. The LADI training dataset is a collection of pictures acquired from an aircraft, whereas the LADI test dataset is a collection of short video clips recorded from a UAV. According to the LADI developers [18, 246], each Human Intelligence Task (HIT) on the MTurk platform asks the human worker if any of the labels in each of the coarse categories are accurate—each HIT only asks about one category at a time. Consequently, each HIT is assigned to three workers to reach an agreement on the label quality. If further validation was required, the HIT was outsourced to two more workers, for five workers per category and image.

With only about 6% to 7% of the 500k images in the LADI training dataset being labeled by human workers, we tackle several challenges that arise from working with a highly-imbalanced and noisy dataset to train a reliable model. The dataset's class imbalance and label noise are illustrated in Figure 6.6. The positive to negative (PN) ratios of the 31 target features calculated from LADI illustrate how the training dataset varies from mostly severely imbalanced (low PN ratios) to reasonably

balanced (high PN ratios) representations. We further compute the Krippendorff's alpha (α) [247], a well-known reliability metric used to measure inter-rater agreement for the annotated labeled training dataset. Unlike other reliability techniques, α handles missing data and is flexible in sample size, category, and the number of workers. The α coefficient is calculated following LADI's described annotation procedure, in which each worker will have the chance to determine whether an image contains a target feature or not. If an annotator never comes across a specific image, this is a missing value. The maximum value for each target feature's α coefficient can only reach $\alpha = 0.51$, indicating the need to correct the label noise and augment the training dataset.

The LADI training dataset was further expanded by including web-crawled images for those very underrepresented features such as water-tower, utility-lines, communication tower, snow/ice, rock, landslide, and smoke/fire. Less than 1000 images for each feature were added from web crawling to improve these features and avoid adding more noise to the data. While crawling for new pictures helps retrieve more relevant samples, the web-crawled pictures may add extra noise to the training data if not utilized properly. Additionally, it is challenging to acquire high-quality low-altitude pictures from the image engine for many of these target features, considering that people seldom take photos from this viewpoint. Thus the web-crawled pictures were not used to balance the training dataset, and the impact of the additional crawled images on the PN ratio is shown in Figure 6.6. Data augmentation methods were used on the training data to improve the model performance, especially for the minority classes. Specifically, the applied augmentation methods include horizontal and vertical flipping with 0.5 probability, 90° rotations with 0.1 probability, contrast change by a factor between 0 and 0.25, and the horizontal and vertical shift within $\pm 10\%$ of the width and height, respectively.

Competing Methods

To ensure that the suggested method is effective, we compare it to a baseline and several competing methods, which are listed below.

- **DCCA [170]**: The deep canonical correlation analysis employs neural networks to exploit the non-linear transformations and learns the representations of images and texts that maximize their correlations.
- **DCE [248]**: The deep collaborative embedding employs a weak supervision technique for refining initial tags and assigning tags to new images via discovering the unified latent space for images and tags.
- **SHIELD [234]**: Experimented with various CNN combinations on the LADI dataset and extended the LADI training data by labeling an unlabeled subset from LADI using Amazon Mechanical Turk.
- **VCL [179]**: Performed a series of experiments to evaluate the roles that objects play in scene comprehension, utilizing various methods for integrating the local-level information (e.g., objects, entities).
- **Ours-InceptionV3-base**: The baseline model consists of five categorical models based on Inception-V3 but, different from the proposed approach, it is trained solely on the soft-labels generated from the human annotators.

Feature Score Model

The feature score model consists of five categorical models based on the InceptionV3 architecture, with each model being trained on the feature scores of a particular category. The weights of these models are pre-trained on ImageNet and then fine-tuned on the disaster-related dataset following the transfer learning process. The last classification head of the network replaces a dense layer implementing the sigmoid

activation function for multi-class soft-label classification. The binary cross-entropy function measures the model loss during training and adjusts the model weights accordingly. Separating each model by category gives more flexibility and alleviates the high class-imbalance problem in the data.

The LADI data is randomly split into two parts: 80% for training and 20% for validation. Each model is trained on small batch sizes containing 16 sample images from the training set. At the end of each epoch, the model's performance is evaluated on the validation set—only the best model with the lowest validation loss is kept. With an initial learning rate of $\eta = 1e{-}4$, the Adam [217] solver is employed to fine-tune the model weights. Models are trained for 100 epochs.

Inference and Ranking

The LADI test dataset is composed of 41 original videos segmented into 1,825 short video clips ranging between two to twenty seconds. Unlike the training dataset, the test set is composed majorly of drone footage. Nonetheless, our methods prove successful in generalizing well across the different tools used to capture the low-altitude images. During inference, the test video shots are split into multiple unique keyframes and fed to the five categorical models to obtain the scores for the 31 features. In order to facilitate content-based retrieval, a shot-level aggregation of the keyframe-level scores is then introduced to rate the video shot according to its significance.

Results and Discussions

Quantitative Results: For each run, the mean average precision (MAP) across 1000 retrieved shots is determined as a measure of the accuracy in identifying the most relevant features in a shot. For a fair comparison among other methods tested

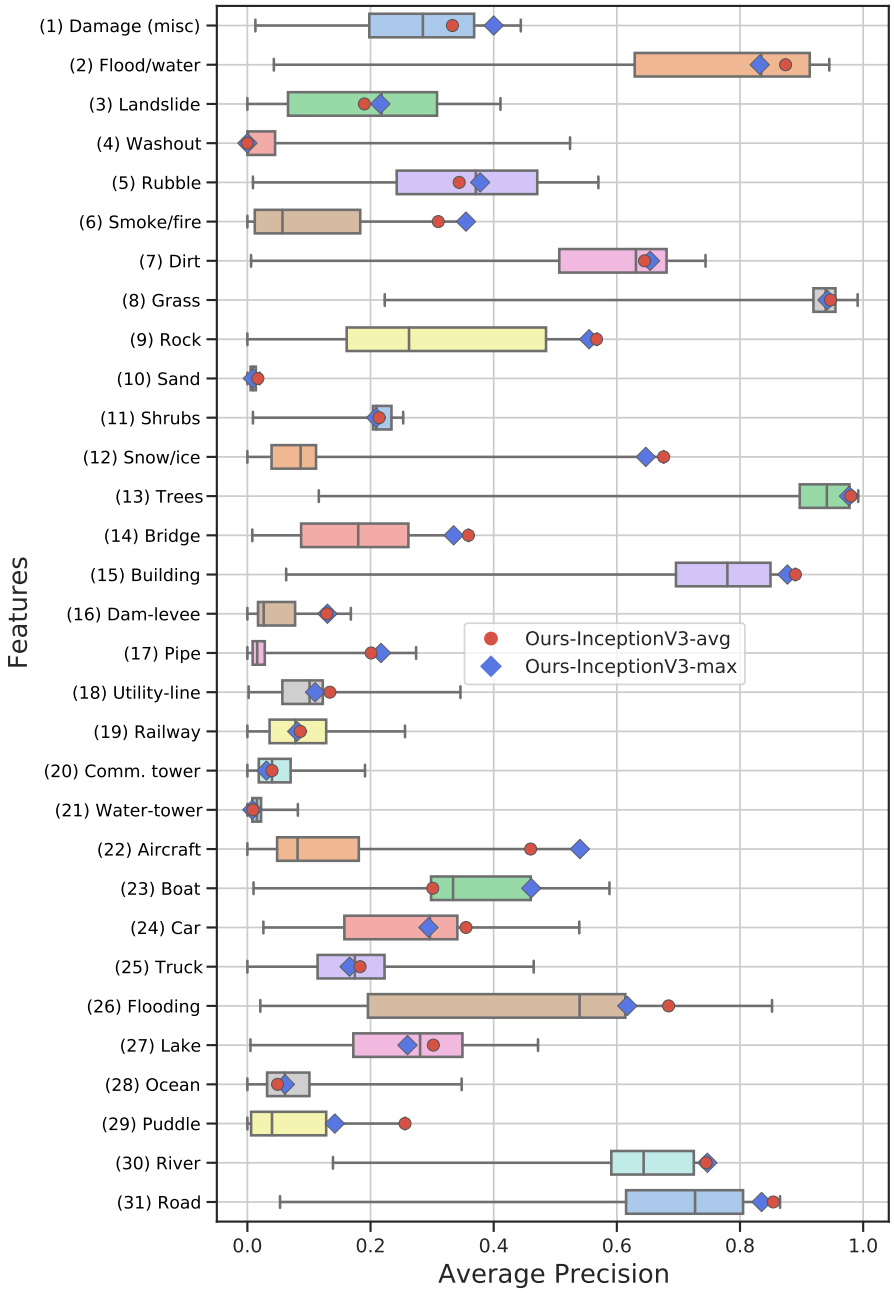


Figure 6.7: Comparison of the boxplot distribution for feature's precision score among all submissions to TRECVID2020-DSDI regardless of the track. The interquartile range of the boxplot is from 25th to 75th percentile. The red dot indicates the placement of our best run among all the submissions. The blue diamond indicates our second-best run.

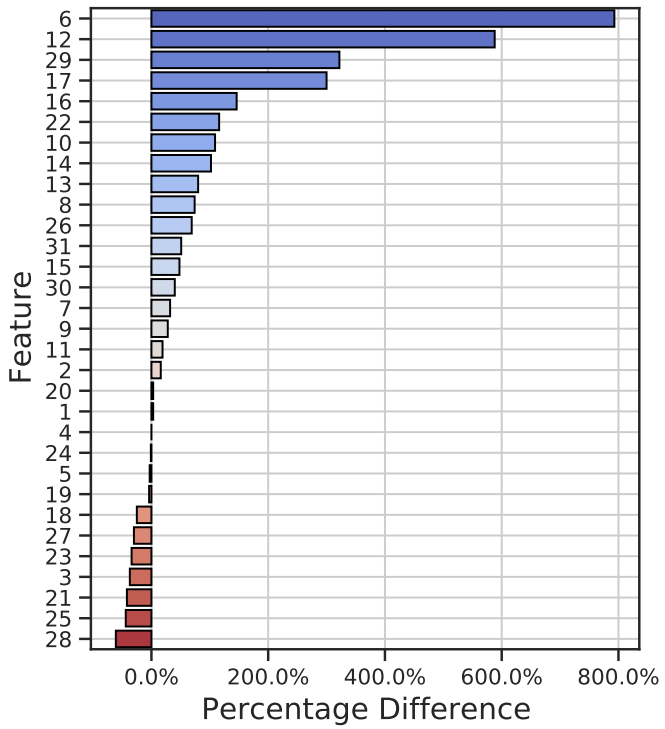


Figure 6.8: Percentage difference between each feature's average precision values from both the baseline and our proposed method. The feature IDs are aligned with those in Figure 6.7.

on the LADI dataset, variants of the proposed framework are compared to the submission of the LADI + Others (O) track—where "Others" in our proposed approach involves the inclusion of data obtained from the web-crawler along with the LADI data to improve the performance of some of the most underrepresented features. A summary of the result comparison is demonstrated in Table 6.5. The proposed method significantly outperforms other tested methods under the same training type. It is worth mentioning that our best-performed approach ranks first among all the solutions in the TRECVID2020-DSDI competition, regardless of the training type.

Furthermore, the average precision (AP) per feature is summarized in Figure 6.7. A boxplot is used to visualize the distribution of the feature-level performance across

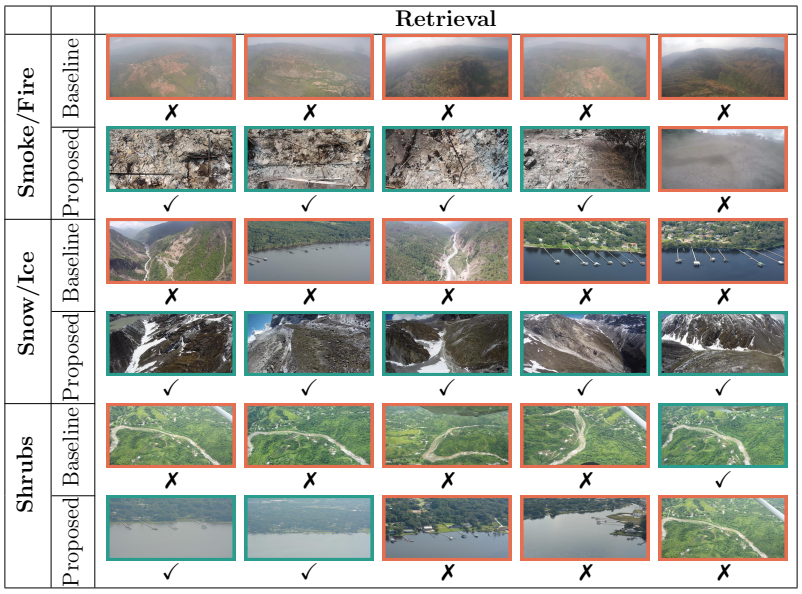
Table 6.5: Comparing the mean precision at 10, 100, and 1000 precision depths, along with the mean average precision (MAP) of our suggested methodology, of our proposed approach to various competing methods and a baseline.

Method	P@10	P@100	P@1000	MAP
DCCA [170]	0.177	0.196	0.210	0.167
DCE [248]	0.329	0.282	0.238	0.205
SHIELD [234]	0.506	0.379	0.236	0.297
VCL [179]	0.232	0.218	0.225	0.176
	0.400	0.346	0.260	0.275
	0.355	0.369	0.264	0.285
	0.471	0.394	0.272	0.333
Ours-InceptionV3-base	0.445	0.404	0.274	0.283
Ours-InceptionV3-top	0.568	0.446	0.278	0.388
Ours-InceptionV3-max	**0.580**	0.444	0.279	0.390
Ours-InceptionV3-avg	0.561	**0.460**	**0.281**	**0.391**

all competition entries, independent of the training type. The red dot and blue diamond represent our best and our second-best submissions, respectively. The comparisons to the feature-level measurements reveal that our proposed method excels in snow/ice, bridge, building, road, and puddle features. We attribute the great performance of most of these features to the effective exploitation of the contextual information derived from the image's metadata. For instance, infrastructure locations such as roads and buildings are well-documented in OSM. Our proposed method was able to effectively leverage to refine and enhance the soft-labels used to train the model to recognize these types of target features.

The proposed approach is also compared to a baseline model. It is noteworthy that the baseline method already achieves a comparable performance to other proposed techniques, confirming that the proposed technique to calculate soft-labels from human workers' annotations effectively reduces some of the noise in the feature scores. By applying the proposed feature fusion with label propagation, we can

Table 6.6: Qualitative results of the first five video clips retrieved by the baseline and the proposed method. Below each video's screenshot, the check (✓) indicates its relevance to the feature; while the cross (✗) marks those videos that have been incorrectly retrieved as false positives.



see the improvements made compared to the final score and each feature's score as shown in Figure 6.8. Very significant improvements can be observed for most of the features. More notably, smoke/fire and snow/ice demonstrated 792% and 587% improvements. By the proposed methods, time and location data were effectively used in instances where labels were extremely limited. Certain features, including landslide and lake, performed worse due to the overlap and ambiguity in some feature definitions.

Qualitative Results: Table 6.6 illustrates the qualitative results from the top five videos retrieved by the baseline and the proposed method for some features, where the video's keyframe that achieves the highest score for that particular feature, is displayed. Because the LADI's target features are so broad, there are many variations in what may be regarded as valid observations within each feature target. Moreover, many of the features have vague meanings that may overlap. The false

positives obtained from the retrieval further demonstrate the ambiguity and broadness that are present in some of the features' meanings and how these limitations have affected the results. As can be seen from this table, for the smoke/fire feature, the baseline method retrieved many examples of environments surrounded by 'fog,' which means the model might have identified some characteristics in 'fog' to be very similar to 'smoke.' However, we also observe that the proposed approach's 5th retrieved observation is a false positive, as the model most likely misconstrues the feature of 'dust' for smoke. Despite the ambiguity and broadness of the feature, the proposed method significantly improves the retrieval performance for smoke/fire by incorporating the historical data of the relevant real-life events. The snow/ice is another feature that benefits from matching the historical data in NOAA's CDO database using time and location from the training image's metadata as queries. The uncertainty and overlap in these feature definitions may be seen more clearly in the case of the feature shrub. A shrub is a kind of foliage that might be difficult to tell apart from a tree from afar. Although our proposed framework effectively retrieves more relevant videos with shrubs, some of the videos categorized as not relevant may include shrubs, or it is not easy to discern.

6.3 Conclusion

This chapter describes a novel approach to identifying damaged buildings from remote sensing images and predicting the level of damage that the building sustained after a disaster event. The proposed model removes the dependency on the available high-quality building footprint geometries by assuming that only an estimate of the damaged building's location and damage level is available to train the model. Weak supervision and perturbations, and noise regularization are critical elements

in developing a robust model to label noise, which can adapt better across multiple domains and various types of areas. The proposed model's predictive results can be used to visualize, identify quickly, flag damaged buildings, and conduct further studies and research. As part of the future work, we will continue to improve the model's performance and develop techniques to better handle the class imbalance in the data by assigning higher weights to the samples from minority classes during training.

Moreover, more than ever, it is now feasible to dispatch a drone ahead of the rescue crew to inspect the impacted region and aid responders to automatically identify those areas that are the most affected and should be prioritized to deliver a timely and appropriate response. The proposed weakly-supervised framework using label propagation aims to predict the chance of a particular catastrophe or environment-related characteristic being present inside a low-altitude snapshot or video. This paper introduces a weakly-supervised deep learning approach developed for automatic disaster scene description of low-altitude pictures captured from an aircraft. The proposed approach is also intended to cut down the time and effort human annotators spend labeling images.

CHAPTER 7

3D ADVANCED VISUALIZATION

7.1 A 3D Virtual Environment for Storm Surge Flooding Animation

This chapter shows an advanced visualization method used in a 3D game engine to provide an immersive and interactive experience created from multi-source data.

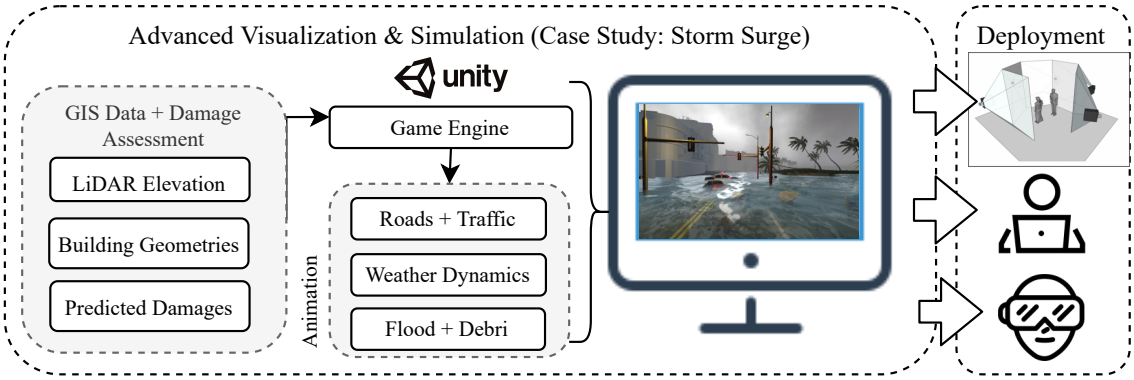


Figure 7.1: Proposed 3D Advanced Visualization & Simulation Platform

Three-dimensional representations of real-world locations have been widely used in computer and mobile phone games and virtual globes to give users unique and immersive experiences. These advanced visualization techniques are widely popular and have been successfully used to analyze complex data interactively. This chapter explicitly explores a case study of a storm-surge simulation developed using the real-life Geographic Information System (GIS) data and the predicted damages to construct this virtual environment automatically.

(a) Street-level view (b) Bird's-eye view

Figure 7.2: The South Beach 3D model is shown from different views during a storm surge.

7.1.1 Motivation

When the wind from a coastal storm, such as a hurricane, pushes water towards the land, it is called a storm surge. Storm surges have disastrous consequences for infrastructure, roadways, and the lives of those who live in flood-prone regions. Officials use storm surges to decide who needs to evacuate. Because most of the state's main cities are located along the coast or in low-lying regions, Florida is particularly vulnerable to storm surges. Since Hurricanes Katrina and Rita in 2005, the state has not been directly impacted by a storm.

On the other hand, Florida has been struck by more direct hurricanes than other states. The greater the population growth in Florida's coastal regions, the greater the danger of a storm claiming thousands of lives. When attempting to depict the regions that may be impacted or have already been affected by a storm, current methods often depend on two-dimensional (2D) visualization. A map of the region with a color-coded scheme to show which regions are most susceptible and which areas are safe is a typical presentation used. Users unfamiliar with the topography of the present area they live in or the hazards of storm surges may be unaware of their degree of risk and make potentially fatal decisions.

7.1.2 System Architecture

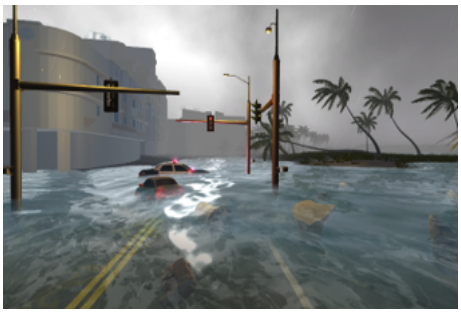
Unity 3D, a sophisticated cross-platform game engine for generating high-end 3D experiences, is used to create the proposed system [16]. We can use Unity to use several pre-built components, eliminating the requirement to start from scratch. For example, the Unity Terrain Engine is renowned for being highly optimized and straightforward to modify. We start by gathering Light Detection and Ranging (LiDAR) data from the National Oceanic and Atmospheric Administration (NOAA) website to create a landscape that resembles the Miami Beach region. LiDAR is a laser-based remote sensing system that measures the time it takes for a sensor to detect reflected light. Compared to other techniques, the advantages of utilizing LiDAR as a method for measuring elevation include its high precision and cost-effectiveness. Now that LiDAR data is becoming more widely available, we can use it to create accurate and dynamic 3D maps. We use the BlenderGIS module to import the LiDAR data into Blender and use the Delaunay triangulation method to build a 3D mesh surface. The surface mesh is imported into Unity. Then, after generating new terrain, we utilize the Object2Terrain module to make the terrain match the mesh's form.

After the landscape has been updated to reflect the various elevation regions of Miami Beach, vegetation and texture have been added to the terrain to make it more realistic. We utilize the information given by Open Street Maps (OSM) to create the buildings and the roadways. We first pick our area of interest (e.g., South Miami Beach) on the OSM website and then download the file that includes all relevant information (e.g., building footprint, building height, roads, and so on) for that area. Then, using OSM2World, we generate 3D buildings and roads, which we then export into Unity as a single file. The components that mimic the storm are added to both the landscape and the structures in the Unity environment. A water

module from Unity's basic components is added to the terrain to simulate the ocean and surging waves. This module is very adaptable. Many characteristics, such as wave frequency, wave steepness, wave speed, and wave direction, may be readily adjusted to meet the requirements of a storm surge animation. Strong winds and torrential rainfall often accompany hurricanes. To make the simulation seem more genuine, our model includes a rain motion and tree movement caused by the winds.

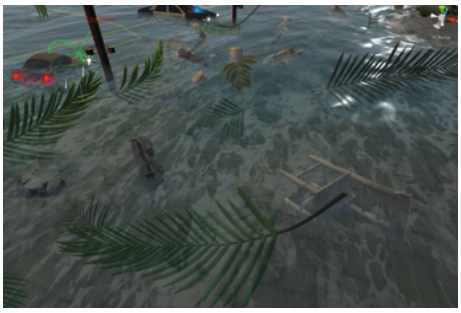
The following is a list of different parameters that the user can change to create different types of storm surge scenarios.

- *Wind Scale*: When the user selects the category of the hurricane, the intensity of the wind is set according to the Saffir-Simpson hurricane wind scale. Unity's built-in wind zone component makes it easy to set the parameters for the changes in the wind.
- *Rain Intensity*: The rain animation was created using the Unity Particle System. A user can change the parameters that affect the visual of the rain. Such parameters include the intensity of the rain and the force factor of the wind on the rain.
- *Wave Intensity*: Waves can flood the land, smash the trees and infrastructures, and carry scattered debris in multiple directions.
- *Debris Properties*: A parameter that a user can set for debris is the average weight which determines how much the scattered pieces can move according to the force of the wind and how much damage it can produce when it hits a building.
- *Tree Bend and Break Factor*: The tree bends according to the effects of the wind. Under certain conditions, the wind can be strong enough to break a branch from a tree.

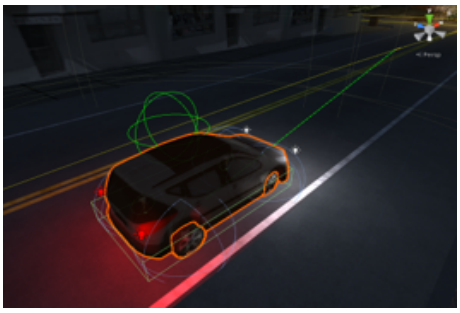


(a) Flood

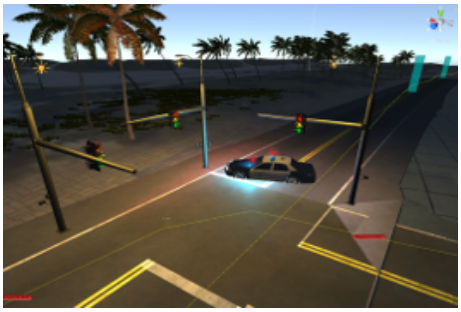
(b) Fog



(c) Debri



(d) Waypoint System



(e) Intersection

(f) Falling Tree

Figure 7.3: Animated features of the proposed 3D visualization.

7.1.3 Demonstration

Figure 7.4: Demonstrating the 3D animation to a group of students at the I-CAVE facility at Florida Internal University.

The animation is built using a 3D game engine, Unity, which opens many possibilities for its deployment, namely, its capability for cross-platform support for the popular operating systems (e.g., Windows, Mac, and Linux). The use of GIS data makes it possible for the model of the city to match a real-life environment accurately. The Light Detection and Ranging (LiDAR) downloaded from The National Oceanic, and Atmospheric Administration (NOAA) website is used to produce the terrain that depicts an accurate visualization of South Miami Beach. LiDAR is a remote sensing technology that produces point clouds, and each point represents

the elevation at a specific location. The bare-earth points can be extracted from the LiDAR point cloud data, and a digital elevation model (DEM) can be created, a grayscale raster where each pixel contains the height information for each location. Unity makes it easy to connect the animation with the Integrated Computer Augmented Virtual Environment (I-CAVE), a visualization and research facility ideal for presenting 3D virtual environments.

I-CAVE is a virtual reality experience that consists of five 9×5 foot high-resolution displays placed in a hexagonal layout with a surround sound system. The devices include hardware and software tracking capabilities that monitor movement in space and synchronize with projected visuals to immerse users in the simulated world. I-CAVE gives users the capability of navigating through the terrain as an immersive experience.

We demonstrate how the wind pushes the seawater towards the city during a hurricane, causing the streets to flood. The user has two choices for navigation. The bird's-eye view (as shown in Figure 1) shows an overworld perspective. Having a top-level view of the overall scene gives the user a broad sense of how the storm surge affects the city. As can be seen in the figure, waves are surging towards the land and mainly affecting the buildings that sit closer to the coast. Another form of navigation is the human-scale point of view (as shown in Figure 2), where the user can walk around and observe more specific affected areas. The human-scale point-of-view is integrated with I-CAVE for an immersive experience. I-CAVE is also ideal for presenting 3D virtual environments to a group of people as shown in Figure 7.4.

7.2 Conclusion

We examine a storm surge simulation developed using a mix of real-world Geographic Information System (GIS) data and predicted damages to build an automated virtual environment. In particular, we use a 3D game engine to demonstrate advanced visualization techniques to create an immersive and interactive environment. The incorporation of GIS data allows the city model to reflect the real-world surroundings correctly. We demonstrate in the virtual environment how the wind carries seawater towards the city during a storm, causing the streets to flood. The 3D environment is integrated with I-CAVE for an immersive experience. Our goal is to make the experience of learning about the repercussions of natural disasters more engaging.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

8.1 Conclusions

A comprehensive framework for intelligent data analytics utilizing deep learning for data science is presented in this dissertation. We designed the intelligent data analytics framework to adapt to various data sources and modalities and execute the required transformations to prepare the data for fusion and modeling. Our method uses the variety and relevance of additional data from a variety of sources that may assist explain the context information in the input data. Data gathered might include images, tags, and thoughts detected by classifiers trained on broader data standards. The proposed framework consists mainly of the following major components: (1) the fusion of data from multiple sources and modalities, (2) fine-scale pattern modeling using deep feature fusion, (3) weakly supervised training, and (4) the 3D advanced visualization. These components are combined into a single entity to offer novel deep learning-based solutions to current data science problems. The following is a summary of each component:

- Data from many sources and formats provide a wealth of helpful information. This component offers a variety of ways for using knowledge from diverse data modalities and sources via fusion. Specifically, two applications are used to test the proposed approaches. The first application is a DNN trained to simulate particle-based energy systems by combining the predictions from previously established empirical correlations with scientific information. Our proposed method provided more precise predictions and showed an exceptional capacity to adapt to a wide range of experimental settings than past models. Accord-

ing to the RMSE measure, our proposed methods could decrease the error to about 30% compared to the baseline performance. We envision this tool as a reliable and inexpensive industrial-scale design and application technique. The second application employed a multi-source weak supervision fusion technique and was trained on a severely skewed dataset with noisy labels. Overall, the research demonstrates that this framework has the potential to save a substantial amount of time and resources while obtaining remarkable outcomes in the job of accurately describing disaster scenes. Our method outperformed all other submitted solutions in the TRECVID2021 [21] DSDI Challenge according to the MAP score (i.e., 0.359) and regardless of the training data utilized.

- Class overlap or inter-class similarity refers to large data overlap between two or more exceptionally similar classes. We construct local characteristics that automatically direct our proposed fine-level pattern modeling strategy for identifying honey bee subspecies in this component using domain knowledge data. Our proposed strategy achieves an 11% improvement according to the F1-score in identifying different bee subspecies. Moreover, we demonstrate how we can recognize and fuse fine-level patterns from the data at several levels of the deep learning architecture, taking into account both local and global interactions between the data's characteristics to output a more accurate prediction. Our solution trained on the single convolutional model and applied to our single model localization methodology obtains an equivalent performance (i.e., xView2 Score=0.8128) as a previously proposed ensemble method (i.e., 0.8119). Our proposed method allowed the prediction ability of two trained models (location and damage assessment) to match an ensemble strategy comprised of eight semantic segmentation models. This compari-

son demonstrates that, compared to semantic segmentation techniques, our methodology can produce state-of-the-art results. Moreover, considering our method is an instance-based approach, it emphasizes identifying the damage to individual buildings and produces consistent predictions that are easier to interpret.

- The rate at which data is gathered far outpaces the rate at which well-curated expert label sets are required to train a successful model using current approaches. We proposed employing a weakly-supervised model, assuming that the label collection is limited and may include mislabeling and errors. Random additive zero-centered Gaussian noise is introduced throughout the training process to induce synthetic perturbations in the target label, supplement the data, and avoid over-fitting. These tiny perturbations must be added to train a robust model to the noise seen in real-world data, such as coordinated position errors and crowd-sourced mislabeling or bias. The model has successfully, and through implicit means, learned to detect different damage levels sustained by buildings, as shown through quantitative and qualitative means. Our component further demonstrates a deep neural network training using low-altitude photography with extremely imbalanced and noisy crowd-sourced labels. We next leverage the images' rich spatio-temporal data and sequence information to improve the training performance of the model via label propagation. The label propagation strategy improves the model's contextual awareness by giving previously unlabeled training data labels and effectively training a model on a large-scale dataset of 500k photos, with just about 7% of the images annotated by a human. As one of the submitted solutions for the TRECVID2020 [20] DSDI Challenge, the suggested approach is tested using

the LADI dataset. Our suggested solution received the highest score among all the participants according to the MAP score (i.e., 0.391).

- We proposed a storm surge simulation built using real-world Geographic Information System (GIS) data and estimated damages to create a virtual environment automatically. Specifically, we employ a 3D game engine to illustrate sophisticated visualization methods to create an immersive and interactive environment. The integration of GIS data enables the city model to represent its actual surroundings accurately. We illustrate in the virtual environment how, during a storm, the wind transports saltwater into the city, causing the streets to flood. I-CAVE converts the proposed 3D environment into an immersive and educational experience.

8.2 Future Work

Earlier chapters demonstrated the effectiveness and efficiency of the proposed framework (shown in Figure 3.1) for intelligent data analytics utilizing deep learning for data science. However, as described below, many challenges will be addressed in future work to enhance the proposed intelligent data analytics framework further.

8.2.1 Weakly Supervised Training with Knowledge Distillation

Large annotated datasets are usually required for modern deep learning techniques; however, gathering such datasets is time-consuming and costly. Fusion seeks to combine information from many modalities and sources to improve a model's predictive accuracy. The extensive data variations observed in these multi-source and

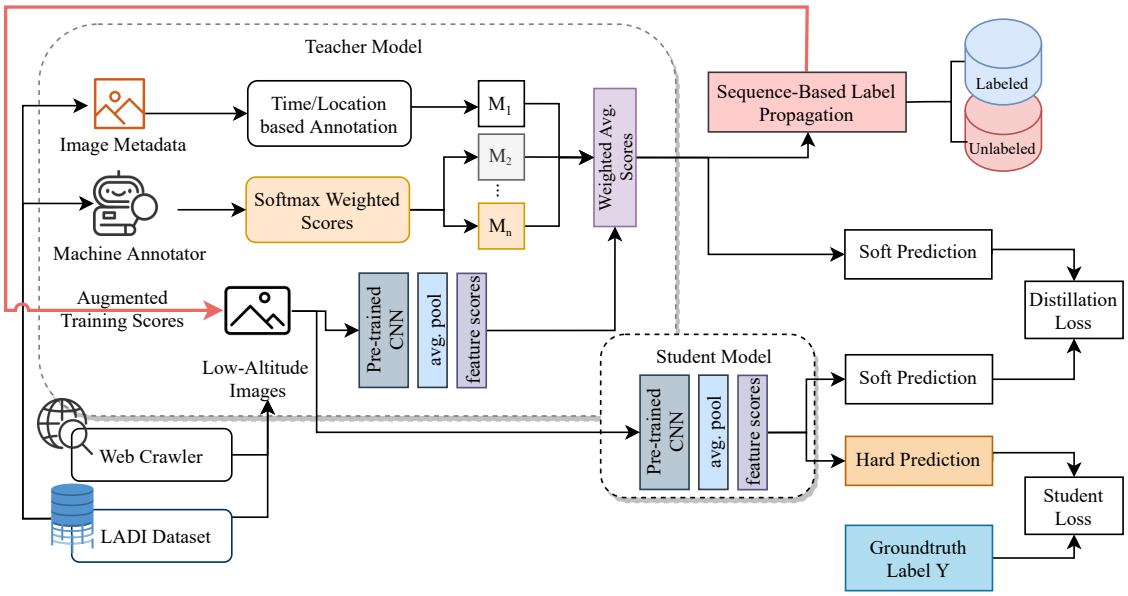


Figure 8.1: Enhanced Weakly Supervised Training with Knowledge Distillation.

multi-modal datasets, on the other hand, provide significant technical challenges for machine learning and deep learning algorithms. To address this problem, researchers developed several fusion methods that combine data from many modalities, such as image, text, and video, to improve the model's dependability, robustness, and accuracy. Developing supervised learning algorithms capable of successfully training datasets with noisy or restricted labels is critical. In the future study, we'll better understand how to effectively utilize the various modalities using a weakly supervised approach based on knowledge distillation, even when their availability during training and testing periods varies.

To further improve the label propagation technique and avoid data leakage, the future work for this component will develop a novel method loosely based on a modified knowledge distillation technique [249] (illustrated in Figure 8.1). Knowledge distillation leverages the big pre-trained (teacher) model on the fusion of different multi-modalities that will then train a more minor (student) model to match

the teacher's performance but have access to a limited set of modality data. The teacher model is assumed to have more information about an image regarding data collected from its geo-tag references, crowd-sourced annotations, predictions from pre-trained classifiers, etc. By minimizing a loss function aiming at matching softening teacher logits and ground-truth labels, knowledge is transmitted from the teacher model to the student model. Knowledge distillation has been proposed as a model compression technique in which a small model is taught to resemble a more prominent, pre-trained model (or ensemble of models). This training environment is often teacher-student, with the big model serving as the teacher and the small model serving as the student.

8.2.2 Advanced Synthetic Data Generation

Figure 8.2: Preliminary results of the potential synthetic data generation tool trained on the building damage assessment dataset.

Existing deep learning algorithms necessitate massive datasets with enough variability to reflect a wide range of real-world events and circumstances. However, many data science applications may not have access to high quantities of data. Au-

tomatically producing synthetic data for deep learning model training is an effective solution. Recent advances in deep learning enable the transformation of data between different domains [230]. For instance, by integrating the primary attributes in the content images with the artistic style (i.e., color, visual structures, etc.), it is feasible to generate minority-class samples. As illustrated in Figure 8.2, the building damage assessment dataset known as xBD [17] can significantly benefit from an advanced synthetic data generator. Networks such as the StyleGAN2 [250] can help generate more damage-based building samples from different disaster types (i.e., flood, wind, etc.).

8.2.3 Capturing Data Structural Knowledge for Deep Feature Fusion

In this dissertation, deep feature fusion in a fine-level pattern modeling method is proposed to offer a robust representation capacity for identifying easily confused inter-class data samples. This preliminary model shows that the latest techniques for finding AI keypoints in geometric morphometrics can be successfully employed. The preliminary model starts by identifying morphologically distinct features from raw images, specifically the previously introduced 19 landmarks. Experts in apiology have closely studied and identified these key landmarks that can help categorize honey bees into distinct subspecies. Researchers can access data from various geographic locations, determine population changes, refine species identification tools, speed up the detection of evolutionary changes, and increase data accessibility by digitizing and integrating geomorphometric, genetic, behavioral, and environmental data for single individual insect specimens.

As previously demonstrated in Section 5.1, the similarity among the wing images' key landmarks makes it difficult to identify the correct keypoint. We observed that the errors in the predicted keypoint locations are primarily due to the confusion and the similarity of different keypoint locations. Taking advantage of the wing's keypoint structure, we proposed to extend.

The difference in Euclidean distance for each keypoint is simply added to calculate Euclidean loss as follows:

$$\mathcal{L}_{Structure} = \sum_{i} \sum_{j} Euclidean(k_{i_{true}} - k_{j_{true}}, k_{i_{pred}} - k_{j_{pred}}) * A_{i,j} \tag{8.1}$$

where k_i is the ith keypoint and $A_{i,j} \in [0, 1]$ is an indicator function that k_j is adjacent to k_i. In other words, we want the difference between each pair of adjacent keypoints to match the actual difference by some distance measure. This loss function helps guide the network to produce keypoints that maintain the linear geometry of the target keypoints by minimizing the difference.

BIBLIOGRAPHY

[1] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M. L. Shyu, S. C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," aug 2018. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3234150

[2] H. Tian, M. Presa-Reyes, Y. Tao, T. Wang, S. Pouyanfar, M. Alonso Jr., S. Luis, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, "Data Analytics for Air Travel Data: A Survey and New Perspectives," *ACM Computing Surveys*, in press, 2021.

[3] J. D. Eisenberg, D. Banisakher, M. Presa, K. Unthank, M. A. Finlayson, R. Price, and S.-C. Chen, "Toward semantic search for the biogeochemical literature," in *IEEE 18th International Conference on Information Reuse and Integration*. IEEE, 2017, pp. 517–525.

[4] D. M. Banisakher, M. E. Reyes, J. Allen, J. D. Eisenberg, M. A. Finlayson, R. Price, and S. C. Chen, "Ontology-based supervised concept learning for the biogeochemical literature," in *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, 2018, pp. 402–410.

[5] Bernard Marr, "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read," 2018. [Online]. Available: https://www.forbes.com/

[6] M. Presa-Reyes, B. Bogosian, B. Schonhoff, C. Jerauld, C. Moreyra, P. Gardinali, and S. C. Chen, "A Water Quality Research Platform for the Near-real-Time Buoy Sensor Data," in *Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020*, 2020, pp. 287–294.

[7] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 4, pp. 549–553, 2017.

[8] M. E. Presa-Reyes, S. Pouyanfar, H. C. Zheng, H.-Y. Ha, and S.-C. Chen, "Multimedia data management for disaster situation awareness," in *International Symposium on Sensor Networks, Systems and Security*. Springer, 2017, pp. 137–146.

[9] M. E. Presa-Reyes and S.-C. Chen, "Assessing Building Damage by Learning the Deep Feature Correspondence of before and after Aerial Images," in *Proceedings - 3rd International Conference on Multimedia Information Processing and Retrieval, MIPR 2020*, 2020, pp. 43–48.

[10] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura, "Damage detection from aerial images via convolutional neural networks," in *2017 Fifteenth IAPR international conference on machine vision applications*. IEEE, 2017, pp. 5–8.

[11] C. El Morr and H. Ali-Hassan, "Descriptive, Predictive, and Prescriptive Analytics," in *Analytics in Healthcare*. Springer, Cham, 2019, pp. 31–55. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-04506-7_3

[12] T. Bustamante, S. Fuchs, B. Grünewald, and J. D. Ellis, "A geometric morphometric method and web application for identifying honey bee species (apis spp.) using only forewings," *Apidologie*, pp. 1–10, 2021.

[13] O. Ozdemir, R. L. Russell, and A. A. Berlin, "A 3d probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose ct scans," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1419–1429, 2019.

[14] J. L. Crossingham, J. Jenkinson, N. Woolridge, S. Gallinger, G. A. Tait, and C.-A. E. Moulton, "Interpreting three-dimensional structures from two-dimensional images: a web-based interactive 3d teaching model of surgical liver anatomy," *HPB*, vol. 11, no. 6, pp. 523–528, 2009.

[15] M. Presa-Reyes and S.-C. Chen, "Weakly-Supervised Damaged Building Localization and Assessment with Noise Regularization," in *IEEE 4th International Conference on Multimedia Information Processing and Retrieval*. IEEE, 2021, pp. 8–14.

[16] M. E. Presa-Reyes and S.-C. Chen, "A 3D virtual environment for storm surge flooding animation," in *The Third IEEE International Conference on Multimedia Big Data*. IEEE, 2017, pp. 244–245.

[17] R. Gupta, B. Goodman, N. N. Patel, R. Hosfelt, S. Sajeev, E. T. Heim, J. Doshi, K. Lucas, H. Choset, and M. E. Gaston, "Creating xbd: A dataset for assessing building damage from satellite imagery," *ArXiv*, vol. abs/1911.09296, 2019.

[18] J. Liu, D. Strohschein, S. Samsi, and A. Weinert, "Large scale organization and inference of an imagery dataset for public safety," in *2019 IEEE High Performance Extreme Computing Conference, HPEC 2019*, 2019. [Online]. Available: https://github.com/cvondrick/vatic

[19] M. Presa-Reyes, Y. Tao, S.-C. Chen, and M.-L. Shyu, "Deep learning with weak supervision for disaster scene description in low-altitude imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[20] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot, "TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains," in *TRECVID*. NIST, USA, 2020.

[21] G. Awad, A. A. Butt, K. Curtis, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, G. J. F. Jones, and G. Quénot, "Evaluating multiple video understanding and retrieval tasks at trecvid 2021," in *Proceedings of TRECVID 2021*. NIST, USA, 2021.

[22] Y. Wang, "Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 1s, pp. 1–25, 2021.

[23] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[24] Y. Wei, D. Wu, and J. Terpenny, "Decision-level data fusion in quality control and predictive maintenance," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 1, pp. 184–194, 2020.

[25] I. Souissi, N. B. Azzouna, and L. B. Said, "A multi-level study of information trust models in wsn-assisted iot," *Computer Networks*, vol. 151, pp. 12–30, 2019.

[26] J. A. Balazs and J. D. Velásquez, "Opinion mining and information fusion: a survey," *Information Fusion*, vol. 27, pp. 95–110, 2016.

[27] K. Sambhoos, J. Llinas, and E. Little, "Graphical methods for real-time fusion and estimation with soft message data," in *11th International Conference on Information Fusion*. IEEE, 2008, pp. 1–8.

[28] P. Luo and Z. S. Li, "A review of internet of things (iot) based engineering applications and data fusion challenges for multi-rate multi-sensor systems," in *IEEE International Conference on Prognostics and Health Management*. IEEE, 2020, pp. 1–7.

[29] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, "A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare," *Inf. Fusion*, vol. 55, pp. 105–115, 2020.

[30] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, and S.-C. Chen, "Applying data mining techniques to address disaster information management challenges on mobile devices," in *The 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 283–291.

[31] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Inf. Fusion*, vol. 53, pp. 80–87, 2020.

[32] O. Nafea, W. Abdul, M. Ghulam, and M. Alsulaiman, "Sensor-based human activity recognition with spatio-temporal deep learning," *Sensors (Basel, Switzerland)*, vol. 21, 2021.

[33] R. Mondal, D. Mukherjee, P. Singh, V. Bhateja, and R. Sarkar, "A new framework for smartphone sensor-based human activity recognition using graph neural network," *IEEE Sensors Journal*, vol. 21, pp. 11 461–11 468, 2021.

[34] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[35] G. A. Gross, R. Nagi, K. Sambhoos, D. R. Schlegel, S. C. Shapiro, and G. Tauer, "Towards hard+soft data fusion: Processing architecture and implementation for the joint fusion and analysis of hard and soft intelligence data," in *2012 15th International Conference on Information Fusion*, 2012, pp. 955–962.

[36] K. Date, G. A. Gross, S. Khopkar, R. Nagi, and K. Sambhoos, "Data association and graph analytical processing of hard and soft intelligence data," in *16th International Conference on Information Fusion*. IEEE, 2013, pp. 404–411.

[37] J. Deng, S. Bei, S. Shaojing, and Z. Zhen, "Feature fusion methods in deep-learning generic object detection: A survey," in *IEEE 9th Joint International*

Information Technology and Artificial Intelligence Conference, vol. 9. IEEE, 2020, pp. 431–437.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[40] Y. Que and H. J. Lee, "Densely connected convolutional networks for multi-exposure fusion," in *International Conference on Computational Science and Computational Intelligence*. IEEE, 2018, pp. 417–420.

[41] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," *arXiv preprint arXiv:1501.02598*, 2015.

[42] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[43] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[44] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[45] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[46] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[47] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement

learning," in *19th ACM International Conference on Multimodal Interaction*, 2017, pp. 163–171.

[48] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011, vol. 40.

[49] G. Shafer, "Dempster-shafer theory," *Encyclopedia of artificial intelligence*, vol. 1, pp. 330–331, 1992.

[50] S. Pouyanfar and S.-C. Chen, "Automatic video event detection for imbalance data using enhanced ensemble deep learning," *International Journal of Semantic Computing*, vol. 11, no. 01, pp. 85–109, 2017.

[51] X.-S. Wei, J. Wu, and Q. Cui, "Deep learning for fine-grained image analysis: A survey," *ArXiv*, vol. abs/1907.03069, 2019.

[52] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang, "Scene change detection by audio and video clues," in *IEEE International Conference on Multimedia and Expo*, vol. 2. IEEE, 2002, pp. 365–368.

[53] S.-C. Chen, R. L. Kashyap, and A. Ghafoor, *Semantic models for multimedia database searching and browsing*. Springer Science & Business Media, 2006, vol. 21.

[54] L. Kabbai, M. Abdellaoui, and A. Douik, "Image classification by combining local and global features," *The Visual Computer*, vol. 35, no. 5, pp. 679–693, 2019.

[55] J. Fang, Y. Zhou, Y. Yu, and S. Du, "Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1782–1792, 2016.

[56] M. AbdelMaseeh, I. Badreldin, M. F. Abdelkader, and M. El Saban, "Car make and model recognition combining global and local cues," in *21st International Conference on Pattern Recognition*. IEEE, 2012, pp. 910–913.

[57] E. Mwebaze, T. Gebru, A. Frome, S. Nsumba, and J. Tusubira, "icassava 2019 fine-grained visual categorization challenge," *arXiv preprint arXiv:1908.02900*, 2019.

[58] J. Yin, A. Wu, and W.-S. Zheng, "Fine-grained person re-identification," *International journal of computer vision*, vol. 128, no. 6, pp. 1654–1672, 2020.

[59] D. M. Sundaram and A. Loganathan, "A new supervised clustering framework using multi discriminative parts and expectation–maximization approach for a fine-grained animal breed classification (sc-mpem)," *Neural Processing Letters*, vol. 52, no. 1, pp. 727–766, 2020.

[60] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLoS computational biology*, vol. 14, no. 12, p. e1006613, 2018.

[61] T. Berg and P. N. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 955–962.

[62] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 5546–5555.

[63] Y. Wang, J. Choi, V. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1163–1172.

[64] Y. Em, F. Gag, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan, "Incorporating intra-class variance to fine-grained visual recognition," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2017, pp. 1452–1457.

[65] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnns for fine-grained visual recognition," *arXiv: Computer Vision and Pattern Recognition*, 2015.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[67] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1458–1465.

[68] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Com-*

puter Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, 2006, pp. 2169–2178.

[69] R. Augustauskas and A. Lipnickas, "Improved pixel-level pavement-defect segmentation using a deep autoencoder," *Sensors*, vol. 20, no. 9, p. 2557, 2020.

[70] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 08 2017. [Online]. Available: https://doi.org/10.1093/nsr/nwx106

[71] A. J. Ratner, S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *VLDB Endowment. International Conference on Very Large Data Bases*, vol. 11 3, pp. 269–282, 2017.

[72] A. Sedova, A. Stephan, M. Speranskaya, and B. Roth, "Knodle: Modular weakly supervised learning with pytorch," *arXiv preprint arXiv:2104.11557*, 2021.

[73] J. Gonsior, M. Thiele, and W. Lehner, "Weakal: Combining active learning and weak supervision," in *International Conference on Discovery Science*. Springer, 2020, pp. 34–49.

[74] B. Settles, "Active learning literature survey," *Science*, vol. 10, no. 3, pp. 237–304, 1995.

[75] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[76] P. Dutta and S. Saha, "A weak supervision technique with a generative model for improved gene clustering," *IEEE Congress on Evolutionary Computation*, pp. 2521–2528, 2019.

[77] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 7158–7166.

[78] L. Bing, S. Chaudhari, R. C. Wang, and W. Cohen, "Improving distant supervision for information extraction using label propagation through lists," in

Conference on Empirical Methods in Natural Language Processing, 2015, pp. 524–529.

[79] J. Wang, X. Shen, and W. Pan, "On transductive support vector machines," *Contemporary Mathematics*, vol. 443, pp. 7–20, 2007.

[80] J. von Kügelgen, M. Loog, A. Mey, and B. Schölkopf, "Semi-supervised learning, causality, and the conditional cluster assumption," in *UAI*, 2020.

[81] Y. Wang, J. Han, Y. Shen, and H. Xue, "Pointwise manifold regularization for semi-supervised learning," *Frontiers of Computer Science*, vol. 15, 2020.

[82] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, Carnegie Mellon University, 2005. [Online]. Available: http://ezproxy.fiu.edu/login?url=https://www.proquest.com/dissertations-theses/semi-supervised-learning-with-graphs/docview/305006312/se-2?accountid=10901

[83] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS*, 2003.

[84] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003.

[85] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[86] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.

[87] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5070–5079.

[88] Z.-H. Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, vol. 1, 2004.

[89] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[90] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[91] Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[92] H. Zhang, L. Jiang, and W. Xu, "Multiple noisy label distribution propagation for crowdsourcing." in *IJCAI*, 2019, pp. 1473–1479.

[93] M.-K. Xie and S.-J. Huang, "Partial multi-label learning with noisy label identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[94] A. K. Menon, B. Van Rooyen, and N. Natarajan, "Learning from binary labels with instance-dependent noise," *Machine Learning*, vol. 107, no. 8, pp. 1561–1595, 2018.

[95] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," *Advances in neural information processing systems*, vol. 26, pp. 1196–1204, 2013.

[96] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015.

[97] C. Scott, G. Blanchard, and G. Handy, "Classification with asymmetric label noise: Consistency and maximal denoising," in *Conference on learning theory*. PMLR, 2013, pp. 489–511.

[98] H. Noh, T. You, J. Mun, and B. Han, "Regularizing deep neural networks by noise: Its interpretation and optimization," *arXiv preprint arXiv:1710.05179*, 2017.

[99] R. Reed and R. J. MarksII, *Neural smithing: supervised learning in feedforward artificial neural networks*. Mit Press, 1999.

[100] S. Pouyanfar, Y. Tao, H. Tian, S.-C. Chen, and M.-L. Shyu, "Multimodal deep learning based on multiple correspondence analysis for disaster management," *World Wide Web*, vol. 22, no. 5, pp. 1893–1911, 2019.

[101] M. Presa-Reyes, P. Mahyawansi, C.-X. Lin, and S.-C. Chen, "DCC-DNN: A Deep Neural Network Model to Predict the Drag Coefficients of Spherical and Non-Spherical Particles Aided by Empirical Correlations," *Powder Technology*, submitted for publication, 2022.

[102] M. E. Presa-Reyes, Y. Tao, R. Ma, S.-C. Chen, and M.-L. Shyu, "Multi-Source Weak Supervision Fusion for Disaster Scene Recognition in Videos," in *The 5th IEEE International Conference on Multimedia Information Processing and Retrieval*, accepted for publication, 2022.

[103] H. Wetherington, "Hurricane Irma recovery." [Online]. Available: https://www.monroecounty-fl.gov/726/Hurricane-Irma-Recovery

[104] National Oceanic and Atmospheric Administration, "Hurricane IRMA: Emergency response imagery of the surrounding regions." [Online]. Available: https://storms.ngs.noaa.gov/storms/irma/download/metadata.html

[105] Monroe County - GIS, "Hurricane Irma - damage assessment." [Online]. Available: http://monroecounty-fl.maps.arcgis.com/apps/webappviewer/index.html?id=87fc14264bdf43fba1669413381dfe3a

[106] S. Xian, K. Feng, N. Lin, R. Marsooli, D. Chavas, J. Chen, and A. Hatzikyriakou, "Rapid assessment of damaged homes in the Florida Keys after Hurricane Irma," *arXiv preprint arXiv:1801.06596*, 2018.

[107] Federal Emergency Management Agency, "Geoplatform.gov," 2020.

[108] ——, "Damage assessment operations manual," 2016. [Online]. Available: https://www.fema.gov/media-library-data/1558541566358-30e29cac50605aae39af77f7e25a3ff0/Damage_Assessment_Manual_4-5-2016.pdf

[109] P. Mitteroecker and P. Gunz, "Advances in geometric morphometrics," *Evolutionary biology*, vol. 36, no. 2, pp. 235–247, 2009.

[110] A. Oleksa and A. Tofilski, "Wing geometric morphometrics and microsatellite analysis provide similar discrimination of honey bee subspecies," *Apidologie*, vol. 46, no. 1, pp. 49–60, 2015.

[111] J. Baba and P. Komar, "Settling velocities of irregular grains at low reynolds numbers," *Journal of Sedimentary Research*, vol. 51, pp. 121–128, 1981.

[112] J. Chen and J. Li, "Prediction of drag coefficient and ultimate settling velocity for high-density spherical particles in a cylindrical pipe," *Physics of Fluids*, vol. 32, no. 5, 2020.

[113] R. P. Chhabra, L. Agarwal, and N. K. Sinha, "Drag on non-spherical particles: An evaluation of available methods," *Powder Technology*, vol. 101, no. 3, pp. 288–295, 1999.

[114] A. T. Corey, "Influence of shape on the fall velocity of sand grains," Ph.D. dissertation, Colorado A & M College, 1949.

[115] F. Dioguardi, D. Mele, and P. Dellino, "A new one-equation model of fluid drag for irregularly shaped particles valid over a wide range of reynolds number: Aerodynamic drag of irregular particles," *Journal of Geophysical Research*, vol. 123, pp. 144–156, 2018.

[116] D. L. Johnson, D. Leith, and P. C. Reist, "Drag on non-spherical, orthotropic aerosol particles," *Journal of Aerosol Science*, vol. 18, no. 1, pp. 87–97, 1987.

[117] S. Kale, "Characterization of aerodynamic drag force on single particles," West Virginia Univ., Morgantown (USA). Dept. of Mechanical and Aerospace, Tech. Rep., 1987.

[118] P. D. Komar and C. E. Reimers, "Grain Shape Effects on Settling Rates," *The Journal of Geology*, vol. 86, no. 2, pp. 193–209, 1978.

[119] G. V. Madhav and R. P. Chhabra, "Drag on non-spherical particles in viscous fluids," *International Journal of Mineral Processing*, vol. 43, no. 1-2, pp. 15–29, 1995.

[120] J. W. Malaika, "M.s. thesis," Master's thesis, University of Iowa, 1949.

[121] G. Mckay, W. Murphy, and M. Hillis, "Settling characteristics of discs and cylinders," *Chem. Eng. Res. Des.*, 1988.

[122] M. Van Melkebeke, C. Janssen, and S. De Meester, "Characteristics and Sinking Behavior of Typical Microplastics including the Potential Effect of Biofouling: Implications for Remediation," *Environmental Science and Technology*, vol. 54, no. 14, pp. 8668–8680, 2020.

[123] E. Pettyjohn and E. Christiansen, "Effect of particle shape on free-settling rates of isometric particles," *Chem. Eng. Prog.*, vol. 44, pp. 157–172, 01 1948.

[124] A. Riazi and U. Türker, "The drag coefficient and settling velocity of natural sediment particles," *Computational Particle Mechanics*, vol. 6, no. 3, pp. 427–437, 2019. [Online]. Available: https://doi.org/10.1007/s40571-019-00223-6

[125] L. Rong, Z. Zhou, and A. Yu, "Lattice–boltzmann simulation of fluid flow through packed beds of uniform ellipsoids," *Powder Technology*, vol. 285, pp. 146–156, 2015.

[126] J. Schmiedel, "Experimentelle untersuchungen uber die fallbewegung von kugeln and scheiben in reibenden flussigkeiten," *Physik. Z. Bd.*, vol. 29, pp. 593–609, 1928.

[127] E. F. Schulz, R. H. Wilde, and M. L. Albertson, "Influence of shape on the fall velocity of sedimentary particles," Ph.D. dissertation, Colorado State University. Libraries, 1954.

[128] M. K. Sharma and R. P. Chhabra, "An experimental study of free fall of cones in Newtonian and Non-Newtonian media: drag coefficient and wall effects," *Chemical Engineering and Processing*, vol. 30, no. 2, pp. 61–67, 1991.

[129] A. W. Sheaffer, "Drag on modified rectangular prisms," *Journal of Aerosol Science*, vol. 18, no. 1, pp. 11–16, 1987.

[130] D. A. Smith and K. F. Cheung, "Settling Characteristics of Calcareous Sand," *Journal of Hydraulic Engineering*, vol. 129, no. 6, pp. 479–483, 2003.

[131] X. Song, Z. Xu, G. Li, Z. Pang, and Z. Zhu, "A new model for predicting drag coefficient and settling velocity of spherical and non-spherical particle in Newtonian fluid," *Powder Technology*, vol. 321, pp. 242–250, 2017.

[132] W. Squires, "The sedimentation of thin discs," Ph.D. dissertation, Massachusetts Institute of Technology, Department of Chemical Engineering, 1936.

[133] G. E. Stringham, D. B. Simons, and H. P. Guy, *The behavior of large particles falling in quiescent liquids.* US Government Printing Office, 1969.

[134] S. Tran-Cong, M. Gay, and E. E. Michaelides, "Drag coefficients of irregularly shaped particles," *Powder Technology*, vol. 139, no. 1, pp. 21–32, 2004.

[135] J. Wang, H. Qi, and C. You, "Experimental study of sedimentation characteristics of spheroidal particles," *Particuology*, vol. 7, no. 4, pp. 264–268, 2009.

[136] P. D. Weidman and I. A. Lasso, "Stokes drag on hollow cylinders and conglomerates." *Phys. Fluids*, vol. 29, no. 12 , Dec. 1986, pp. 3921–3934, 1986.

[137] R. H. Wilde, "Effect of shape on the fall velocity of gravel sized particles," Ph.D. dissertation, Colorado Agricultural and Mechanical College, 1952.

[138] W. W. Willmarth, N. E. Hawk, and R. L. Harvey, "Steady and unsteady motions and wakes of freely falling disks," *Physics of Fluids*, vol. 7, no. 2, pp. 197–208, 1964.

[139] H. Y. Xie and D. W. Zhang, "Stokes shape factor and its application in the measurement of spherity of non-spherical particles," *Powder Technology*, vol. 114, no. 1-3, pp. 102–105, 2001.

[140] S.-C. Chen and R. L. Kashyap, "A spatio-temporal semantic model for multimedia database systems and multimedia information systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 4, pp. 607–622, 2001.

[141] S.-C. Chen and R. Kashyap, "Temporal and spatial semantic models for multimedia presentations," in *1997 International Symposium on Multimedia Information Processing*, 1997, pp. 441–446.

[142] S.-C. C.-L. Shyu and R. Kashyap, "Augmented transition network as a semantic model for video data," *International Journal of Networking and Information Systems, Special Issue on Video Data*, vol. 3, no. 1, pp. 9–25, 2000.

[143] L. Lin and M.-L. Shyu, "Weighted association rule mining for video semantic detection," *International Journal of Multimedia Data Engineering and Management*, vol. 1, no. 1, pp. 37–54, 2010.

[144] P. A. Cundall and O. D. Strack, "A discrete numerical model for granular assemblies," *geotechnique*, vol. 29, no. 1, pp. 47–65, 1979.

[145] D. Gidaspow and B. Ettehadieh, "Fluidization in two-dimensional beds with a jet. 2. hydrodynamic modeling," *Industrial & Engineering Chemistry Fundamentals*, vol. 22, no. 2, pp. 193–201, 1983.

[146] A. Haider and O. Levenspiel, "Drag coefficient and terminal velocity of spherical and nonspherical particles," *Powder technology*, vol. 58, no. 1, pp. 63–70, 1989.

[147] S.-F. Chien, "Settling velocity of irregularly shaped particles," *SPE Drilling & Completion*, vol. 9, no. 04, pp. 281–289, 1994.

[148] A. Hölzer and M. Sommerfeld, "New simple correlation formula for the drag coefficient of non-spherical particles," *Powder Technology*, vol. 184, no. 3, pp. 361–365, 2008.

[149] L. He, D. K. Tafti, and K. Nagendra, "Evaluation of drag correlations using particle resolved simulations of spheres and ellipsoids in assembly," *Powder Technology*, vol. 313, pp. 332–343, 2017.

[150] B. Bhattacharya, R. K. Price, and D. P. Solomatine, "Machine Learning Approach to Modeling Sediment Transport," *Journal of Hydraulic Engineering*, vol. 133, no. 4, pp. 440–450, 2007.

[151] H. D. Yoon, D. T. Cox, and M. Kim, "Prediction of time-dependent sediment suspension in the surf zone using artificial neural network," *Coastal Engineering*, vol. 71, pp. 78–86, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.coastaleng.2012.08.005

[152] F. Oehler, G. Coco, M. O. Green, and K. R. Bryan, "A data-driven approach to predict suspended-sediment reference concentration under non-breaking waves," *Continental Shelf Research*, vol. 46, pp. 96–106, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.csr.2011.01.015

[153] E. B. Goldstein and G. Coco, "A machine learning approach for the prediction of settling velocity," *Water Resources Research*, vol. 50, no. 4, pp. 3595–3601, 2014.

[154] S. N. Yan, T. Y. Wang, T. Q. Tang, A. X. Ren, and Y. R. He, "Simulation on hydrodynamics of non-spherical particulate system using a drag coefficient correlation based on artificial neural network," *Petroleum Science*, vol. 17, no. 2, pp. 537–555, 2020. [Online]. Available: https://doi.org/10.1007/s12182-019-00411-2

[155] S. Balachandar, W. C. Moore, G. Akiki, and K. Liu, "Toward particle-resolved accuracy in Euler–Lagrange simulations of multiphase flow using

machine learning and pairwise interaction extended point-particle (PIEP) approximation," *Theoretical and Computational Fluid Dynamics*, vol. 34, no. 4, pp. 401–428, 2020. [Online]. Available: https://doi.org/10.1007/s00162-020-00538-8

[156] L. T. Zhu, B. Ouyang, H. Lei, and Z. H. Luo, "Conventional and data-driven modeling of filtered drag, heat transfer, and reaction rate in gas–particle flows," *AIChE Journal*, vol. 67, no. 8, pp. 1–13, 2021.

[157] S. Hwang, J. Pan, and L. S. Fan, "A machine learning-based interaction force model for non-spherical and irregular particles in low Reynolds number incompressible flows," *Powder Technology*, vol. 392, pp. 632–638, 2021. [Online]. Available: https://doi.org/10.1016/j.powtec.2021.07.050

[158] K. Luo, D. Wang, T. Jin, S. Wang, Z. Wang, J. Tan, and J. Fan, "Analysis and development of novel data-driven drag models based on direct numerical simulations of fluidized beds," *Chemical Engineering Science*, vol. 231, p. 116245, 2021. [Online]. Available: https://doi.org/10.1016/j.ces.2020.116245

[159] S. Rushd, M. T. Parvez, M. A. Al-Faiad, and M. M. Islam, "Towards optimal machine learning model for terminal settling velocity," *Powder Technology*, vol. 387, pp. 95–107, 2021. [Online]. Available: https://doi.org/10.1016/j.powtec.2021.04.011

[160] K. Thomsen and G. Michon, "Surface area of an ellipsoid," *Spheroids & scalene ellipsoids*, 2004.

[161] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[162] H. N. Yow, M. J. Pitt, and A. D. Salman, "Drag correlations for particles of regular shape," *Advanced Powder Technology*, vol. 16, no. 4, pp. 363–372, 2005.

[163] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1930–1939.

[164] S. F. Buck, "A method of estimation of missing values in multivariate data suitable for use with an electronic computer," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 22, no. 2, pp. 302–306, 1960.

[165] L. A. Clark and D. Pregibon, "Tree-based models," in *Statistical models in S*. Routledge, 2017, pp. 377–419.

[166] S. Galelli and A. Castelletti, "Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling," *Hydrology and Earth System Sciences*, vol. 17, no. 7, pp. 2669–2684, 2013.

[167] M.-F. R. Lee and T.-W. Chien, "Artificial intelligence and internet of things for robotic disaster response," *International Conference on Advanced Robotics and Intelligent Systems*, pp. 1–6, 2020.

[168] E. Weber, N. Marzo, D. P. Papadopoulos, A. Biswas, A. Lapedriza, F. Ofli, M. Imran, and A. Torralba, "Detecting natural disasters, damage, and incidents in the wild," in *The European Conference on Computer Vision*, August 2020.

[169] J. Tang and Z. Li, "Weakly supervised multimodal hashing for scalable social image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 2730–2741, 2018.

[170] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[171] Y. Li, H. Wang, S. Sun, and B. Buckles, "Integrating multiple deep learning models to classify disaster scene videos," in *2020 IEEE High Performance Extreme Computing Conference*, 2020.

[172] S. Okazaki, Q. Kong, M. Klinkigt, and T. Yoshinaga, "Hitachi at TRECVID DSDI 2020," in *TRECVID*. NIST, USA, 2020.

[173] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[174] A. M. Turk, "Amazon mechanical turk," *Retrieved August*, vol. 17, 2012.

[175] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[176] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[177] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[178] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[179] E. Christakis, S. Demertzis, K. Stavridis, A. Psaltis, A. Dimou, and P. Daras, "Towards low-altitude image analysis: Object-enhanced concept detection," in *TRECVID*. NIST, USA, 2020.

[180] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[181] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *IEEE Fourth International Conference on Semantic Computing*. IEEE, 2010, pp. 462–469.

[182] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 04, pp. 715–734, 2001.

[183] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103–7112.

[184] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[185] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[186] H. Inoue, "Data augmentation by pairing samples for images classification," *CoRR*, vol. abs/1801.02929, 2018. [Online]. Available: http://arxiv.org/abs/1801.02929

[187] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European conference on computer vision*, 2018, pp. 466–481.

[188] "Assessing Irma's destruction from the air: Aerial images available." [Online]. Available: https://www.noaa.gov/news/assessing-irma-s-destruction-from-air-aerial-images-available

[189] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," in *Video data management and information retrieval*. IGI Global, 2005, pp. 217–236.

[190] S.-C. Chen, S. H. Rubin, M.-L. Shyu, and C. Zhang, "A dynamic user concept pattern learning framework for content-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 36, no. 6, pp. 772–783, 2006.

[191] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Effective supervised discretization for classification based on correlation maximization," in *2011 IEEE International Conference on Information Reuse & Integration*. IEEE, 2011, pp. 390–395.

[192] L. Zheng, C. Shen, L. Tang, C. Zeng, T. Li, S. Luis, and S.-C. Chen, "Data mining meets the needs of disaster information management," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 5, pp. 451–464, 2013.

[193] S. Xian, K. Feng, N. Lin, R. Marsooli, D. Chavas, J. Chen, and A. Hatzikyriakou, "Brief communication: Rapid assessment of damaged residential buildings in the Florida Keys after Hurricane Irma," *Natural Hazards and Earth System Sciences*, vol. 18, no. 7, 2018.

[194] T. Ci, Z. Liu, and Y. Wang, "Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression," *Remote Sensing*, vol. 11, no. 23, p. 2858, 2019.

[195] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[196] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *CoRR*, vol. abs/1804.07612, 2018. [Online]. Available: http://arxiv.org/abs/1804.07612

[197] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[198] T. Valentijn, J. Margutti, M. van den Homberg, and J. Laaksonen, "Multi-hazard and spatial transferability of a cnn for automated building damage assessment," *Remote Sensing*, vol. 12, no. 17, p. 2839, 2020.

[199] R. Gupta and M. Shah, "Rescuenet: Joint building segmentation and damage assessment from satellite imagery," *CoRR*, vol. abs/2004.07312, 2020. [Online]. Available: https://arxiv.org/abs/2004.07312

[200] H. Hao, S. Baireddy, E. R. Bartusiak, L. Konz, K. LaTourette, M. Gribbons, M. Chan, M. L. Comer, and E. J. Delp, "An attention-based system for damage assessment using satellite imagery," *arXiv preprint arXiv:2004.06643*, 2020.

[201] Y. Shen, S. Zhu, T. Yang, C. Chen, D. Pan, J. Chen, L. Xiao, and Q. Du, "Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[202] L. Gueguen and R. Hamid, "Large-scale damage detection using satellite imagery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1321–1328.

[203] F. Nex, D. Duarte, F. G. Tonolo, and N. Kerle, "Structural building damage detection with deep learning: Assessment of a state-of-the-art cnn in operational conditions," *Remote sensing*, vol. 11, no. 23, p. 2765, 2019.

[204] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the united states," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2600–2614, 2018.

[205] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[206] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sensing*, vol. 11, no. 3, p. 227, 2019.

[207] H. J. Miller, "Tobler's first law and spatial analysis," *Annals of the association of American geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[208] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, "Masked label prediction: Unified message passing model for semi-supervised classification," *arXiv preprint arXiv:2009.03509*, 2020.

[209] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen *et al.*, "Dynamic sampling in convolutional neural networks for imbalanced data classification," in *IEEE conference on multimedia information processing and retrieval*. IEEE, 2018, pp. 112–117.

[210] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[211] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[212] D. xView, "1st place solution for xview2: Assess building damage challenge." [Online]. Available: https://github.com/DIUx-xView/xView2_first_place

[213] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[214] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.

[215] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[216] X. Jia, S. Song, W. He, Y. Wang, H. Rong, F. Zhou, L. Xie, Z. Guo, Y. Yang, L. Yu *et al.*, "Highly scalable deep learning training system

with mixed-precision: Training imagenet in four minutes," *arXiv preprint arXiv:1807.11205*, 2018.

[217] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[218] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[219] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[220] K. Zhang, J. Yan, and S.-C. Chen, "Automatic construction of building footprints from airborne lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2523–2533, 2006.

[221] J. Yan, K. Zhang, C. Zhang, S.-C. Chen, and G. Narasimhan, "Automatic construction of 3-D building model from airborne lidar data through 2-d snake algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 3–14, 2014.

[222] K. Zhang, S.-C. Chen, D. Whitman, M.-L. Shyu, J. Yan, and C. Zhang, "A progressive morphological filter for removing nonground measurements from airborne lidar data," *IEEE transactions on geoscience and remote sensing*, vol. 41, no. 4, pp. 872–882, 2003.

[223] L. Zheng, C. Shen, L. Tang, T. Li, S. Luis, S.-C. Chen, and V. Hristidis, "Using data mining techniques to address critical information exchange needs in disaster affected public-private networks," in *16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 125–134.

[224] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification," in *IEEE international symposium on multimedia*. IEEE, 2015, pp. 483–488.

[225] S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. Iyengar, "Multimedia big data analytics: A survey," *ACM computing surveys*, vol. 51, no. 1, pp. 1–34, 2018.

[226] Y. Shen, S. Zhu, T. Yang, and C. Chen, "Cross-directional feature fusion network for building damage assessment from satellite imagery," *arXiv preprint arXiv:2010.14014*, 2020.

[227] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Readings in computer vision*, pp. 638–643, 1987.

[228] E. J. Kirkland, "Bilinear interpolation," in *Advanced Computing in Electron Microscopy*. Springer, 2010, pp. 261–263.

[229] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. [Online]. Available: http://code.google.com/p/cuda-convnet/

[230] S. Pouyanfar, Y. Tao, S. Sadiq, H. Tian, Y. Tu, T. Wang, S.-C. Chen, and M.-L. Shyu, "Unconstrained flood event detection using adversarial data augmentation," in *2019 IEEE International Conference on Image Processing*. IEEE, 2019, pp. 155–159.

[231] S. Kim, H. Kim, and Y. Namkoong, "Ordinal classification of imbalanced data with application in emergency and disaster information services," *IEEE Intelligent Systems*, vol. 31, no. 5, pp. 50–56, 2016.

[232] J. Mao, K. Harris, N.-R. Chang, C. Pennell, and Y. Ren, "Train and deploy an image classifier for disaster response," in *2020 IEEE High Performance Extreme Computing Conference*. IEEE, 2020, pp. 1–5.

[233] E. Sava, L. Clemente-Harding, and G. Cervone, "Supervised classification of civil air patrol (CAP)," *Natural hazards*, vol. 86, no. 2, pp. 535–556, 2017.

[234] M. Zaffaroni, F. Oldani, and C. Rossi, "Independent category classifiers for emergency scene description using deep learning approaches," in *TRECVID*. NIST, USA, 2020.

[235] B. Lewis, "Civil air patrol offers local support," *TechBeat Dated*, pp. 10–13, 2014.

[236] "Civil air patrol begins deploying small drones for search and rescue," Nov 2019. [Online]. Available: https://www.airforcemag.com/civil-air-patrol-begins-deploying-small-drones-for-search-and-rescue/

[237] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang, "A unified framework for image database clustering and content-based retrieval," in *The 2nd ACM international Workshop on Multimedia databases*, 2004, pp. 19–27.

[238] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 319–337, 2001.

[239] D. Mulfari, A. Celesti, M. Fazio, M. Villari, and A. Puliafito, "Using google cloud vision in assistive technology scenarios," in *IEEE Symposium on Computers and Communication*. IEEE, 2016, pp. 214–219.

[240] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[241] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[242] C. Bansal, A. Singla, A. K. Singh, H. O. Ahlawat, M. Jain, P. Singh, P. Kumar, R. Saha, S. Taparia, S. Yadav *et al.*, "Characterizing the evolution of indian cities using satellite imagery and open street maps," in *ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 87–96.

[243] I. S. Data, "National climatic data center (ncdc)," *Asheville, NC*, 2001.

[244] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.1212303

[245] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A comparison of semantic similarity methods for maximum human interpretability," in *Artificial Intelligence for Transforming Business and Society*, vol. 1. IEEE, 2019, pp. 1–4.

[246] J. Liu and A. Weinert, "Low Altitude Disaster Imagery (LADI) Dataset," https://github.com/LADI-Dataset/ladi-overview, 2019.

[247] K. Krippendorff, "Computing krippendorff's alpha-reliability," *University of Pennsylvania ScholarlyCommons*, 2011. [Online]. Available: https://repository.upenn.edu/asc_papers/43

[248] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2070–2083, 2018.

[249] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[250] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

VITA

MARIA EUGENIA PRESA-REYES

June 25, 1994	Born, Santiago de Cuba, Cuba
2015	B.Sc., Computer Science Florida International University Miami, Florida
2016	M.S., Computer Science Florida International University Miami, Florida
2017–2022	Ph.D. Candidate in Computer Science Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Maria Presa-Reyes, Pratik Mahyawansi, Beichao Hu, Cheng-Xian Lin, and Shu-Ching Chen, "DCC-DNN: A Deep Neural Network Model to Predict the Drag Coefficients of Spherical and Non-Spherical Particles Aided by Empirical Correlations," submitted for publication, *Powder Technology*, 2022.

Maria Presa-Reyes, Shu-Ching Chen, "Multi-Source Weak Supervision Fusion for Disaster Scene Recognition in Videos," accepted for publication, *IEEE International Conference on Multimedia Information Processing and Retrieval*, 2022.

Maria Presa-Reyes, Yudong Tao, Erik Coltey, Tianyi Wang, Rui Ma, Shu-Ching Chen, and Mei-Ling Shyu, "Florida International University - University of Miami TRECVID 2021 DSDI Track," In *TRECVID*, NIST, USA, 2021. (Notebook Paper)

Maria Presa-Reyes, Yudong Tao, Shu-Ching Chen, and Mei-Ling Shyu. "Deep Learning With Weak Supervision for Disaster Scene Description in Low-Altitude Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-10.

Haiman Tian, Maria Presa-Reyes, Yudong Tao, Tianyi Wang, Samira Pouyanfar, Miguel Alonso Jr., Steven Luis, Mei-Ling Shyu, Shu-Ching Chen, Sundaraja Sitharama Iyengar, "Data Analytics for Air Travel Data: A Survey and New Perspectives," *ACM Computing Surveys* 54, no. 8 (2021): 1-35.

Maria Presa-Reyes, Shu-Ching Chen, "Weakly-Supervised Damaged Building Localization and Assessment with Noise Regularization," *International Conference on Multimedia Information Processing and Retrieval*, Tokyo, Japan, September 8-10, 2021.

Maria Presa-Reyes, Biayna Bogosian, Bradley Schonhoff, Christopher Jerauld, Christian Moreyra, Piero Gardinali, Shu-Ching Chen, "A Water Quality Research Platform for the Near-real-time Buoy Sensor Data," *The 21st IEEE International Conference on Information Reuse and Integration for Data Science*, Tuscany Hotel, Las Vegas, Nevada, USA, pp. 287-294, August 11-13, 2020.

Maria Presa-Reyes, Shu-Ching Chen, "Assessing Building Damage by Learning the Deep Feature Correspondence of Before and After Aerial Images," *IEEE 3rd International Conference on Multimedia Information Processing and Retrieval*, Shenzhen, Guangdong, China, pp. 43-48, August 6-8, 2020.

Maria Presa-Reyes, Yudong Tao, Shu-Ching Chen, and Mei-Ling Shyu. "Florida International University-University of Miami TRECVID 2020 DSDI Track." In *TRECVID*, NIST, USA, 2020. (Notebook Paper)

Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, S.S. Iyengar, "A Survey on Deep Learning: Algorithms, Techniques, and Applications," *ACM Computing Surveys*, Volume 51 Issue 5, Article No. 92, January 2019.

Maria E. Presa-Reyes, Samira Pouyanfar, Hector Cen Zheng, Hsin-Yu Ha, Shu-Ching Chen, "Multimedia Data Management for Disaster Situation Awareness," *International Symposium on Sensor Networks, Systems and Security*, Lakeland, FL, USA, pp. 137-146, 8/31-9/2, 2017.

Maria Presa-Reyes, Shu-Ching Chen, "A 3D Virtual Environment for Storm Surge Flooding Animation," *The Third IEEE International Conference on Multimedia Big Data*, Laguna Hills, California, USA, pp. 244-245, April 19-21, 2017. (Demo Paper)